An Analysis of Item Mapping and Test Reporting Strategies

Final Report

October 2003

Joseph M. Ryan, Ph.D
Arizona State University West

A research project initiated by SERVE in collaboration with
The South Carolina Department of Education

Funding provided by The National Science Foundation
Award No. REC – 99787977

SERVE
P.O. Box 5367
Greensboro, North Carolina 27435
www.serve.org
1-800-755-3277

Executive Summary

This project was designed to study those features and formats of score reports that make them useful to educators for identifying students' strengths and weakness and for designing, monitoring, and adjusting instructional programs. The project included a review of assessment reporting research literature and an analysis of current assessment reporting practices. Field-based educators who are deeply involved in curriculum, instruction, and assessment activities in their schools and school districts participated in two focus groups to provide insights and suggestions about the substance and formats of various score reporting approaches.

The critical information and features of score reports that might make them especially useful were identified in the first focus group. This information was used to design six score reporting strategies and formats that were reviewed and evaluated by the second focus group. The six score reporting formats designed, developed, and then evaluated in this study are:

1. Item Content Objective Mapping.
2. Achievement Performance Level Narrative.
3. Strand Achievement Level for Individual Students.
4. Strand Achievement Level for Groups.
5. Observed, Expected, and Differences in Strand and Item Performance for a Group.
6. Observed, Expected, and Differences in Strand and Item Performance at the Achievement Level Cut Scores.

The evaluation of the score reporting formats employed qualitative data from the focus group and quantitative ratings provided by focus-group participants.

The review of research and practice and the results of the project study led to consistent advice about maximizing the value of score reports. Numerous suggestions and guidelines are provided. Among the suggestions are the following: score reports should be simple, clear, uncluttered, and concise; print features such as font size, use of bold, etc, are important; jargon and technical language should be avoided; critical information should be highlighted; graphs, charts, and tables should be kept simple and should be explained with text; score information should be related to content standards as explicitly as possible; the finest level of detail that is still reliable should be reported; some form of normative information is useful; and, information about reliability and precision should be provided.

The value of field testing score reports with their intended audiences through focus groups is strongly recommended. Conducting future studies that document what teachers and others actually do with score report information seems like the next step in this line of research.

Acknowledgements

The author would like to acknowledge with appreciation the support of the South Carolina Department of Education, and especially

Theresa Siskind, Director of the Office of Assessment, for the opportunity to conduct this study and for the measurement advice and suggestions

Necati Engec and Joe Saunders for data analysis, preparation of reporting examples and formats, and sound and insightful psychometric advice

Judy Hair for arranging the focus groups used in this study and

The South Carolina Educators, listed in Appendix A, who participated in the two focus groups.

Jeanne Miyasaka, Educational Measurement Systems, for reviewing and critiquing earlier versions of this report and for editing this final version.

This research was supported by the National Science Foundation, Award No. REC – 99787997 granted to SERVE at the University of North Carolina at Greensboro. Wendy McColskey at SERVE was the project director for this NSF award. For other reports generated through this project, contact her at wmccolsk@serve.org.

Table of Contents

List of Tables

List of Figures

Section 1

Introduction and Overview of the Project and the Report

A. Introduction and Section Overview

The purpose of this project is to explore, develop, and evaluate various approaches that can be used to report students' test scores in ways that are as informative and helpful to students, parents, and educators as possible. The educational practices and research activities that provide the context for the current study all have in common the focus on providing substantively based interpretation of students' test scores.

Ultimately the central question is, as it has been for decades, "What do test scores tell us about what students know and can do?" An early assertion that motivates the examination of this question was provided by Flanagan (1951) more than a half a decade ago when he wrote, "Test scores are meaningful and valuable to the extent that they can be interpreted in terms of capacities, abilities and accomplishments of educational significance." The author wrote this while he was trying to understand the meaning of measurement units, scales, and norms.

The project described in this report examines state assessment score reporting in South Carolina. Although certain activities in the project are focused on one state, many of the results seem to be broadly applicable to large-scale assessment programs in general.

The project evolved in South Carolina for a number of reasons. The state had discontinued the reporting of item-response summaries, and some educators indicated an interest in continuing to have this kind of information still available. In general, the Department of Education was receiving increasing numbers of requests for assessment information that could be used to review and guide instruction. Various issues of score reporting and interpretation had been discussed with the South Carolina Technical Advisory Committee (TAC) on numerous occasions.

The South Carolina Department of Education (SCDE) has a strong collaborative relationship with the SouthEastern Regional Vision for Education (SERVE) and SCDE and SERVE entered into a collaborative effort to pursue this project and research. The project was supported by the National Science Foundation through SERVE.

State Assessment Programs

Every state and even many large school districts have some form of large-scale assessment program in place. In some states, students in selected grades take the state standards-based assessment, a norm-referenced assessment, and NAEP assessments.

Large-scale assessment programs are generally proposed and supported for two reasons. First, state testing programs are used for accountability purposes with a variety of data being used to review, rate, and monitor schools and school districts. A second purpose of state assessment programs is to provide information about students' learning that can be used to diagnose their strengths and weaknesses and the effectiveness of various instructional strategies, school curricula, and programs. This information is collected to plan and implement curriculum and instruction that benefits students the most.

The use of testing programs to support instruction and learning assumes that assessment results are useful to educators for these purposes. In a review of research on test score reporting, however, Goodman and Hambleton (2003) noted that many users of assessment data have difficulty interpreting and understanding results presented in large-scale assessment reports (p. 4).

Given the extraordinary expense involved in developing and operating state assessment programs and the promise that such programs will support instruction and learning, it seems critical that research be focused on developing procedures and formats that make score report information as useful as possible to educators. At the same time, it is equally critical that assessment information meet professional measurement standards such as validity, reliability, fairness, and others as described in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education [NCME], 1999).

No Child Left Behind (NCLB)

An additional motivation to examine the efficacy of students score reports can be found in the No Child Left Behind Act of 2001. NCLB requires that individual results must be reported for all students who take part in the annual assessments and states are required to:

> produce individual student interpretive, descriptive, and diagnostic reports…that allow parents, teachers, and principals to understand and address the specific academic needs of students, and include information regarding achievement on academic assessments aligned with State academic achievement standards, and that are provided to parents, teachers, and principals, as soon as is practicably possible after the assessment is given, in an understandable and uniform format, and to the extent practicable, in a language that parents can understand. (NCLB, 2001, § 1111[b][3][C][xii])

The requirements of NCLB certainly set very high expectations for the quality of the information contained in state assessment reports that will be provided to parents, teachers, and principals. The expectation that parents, teachers, and principals can use assessment results to understand specific academic needs of students assumes assessment reporting

procedures and formats that are highly informative and effective in communicating their information.

B. Project Activities

This project was designed to study the features and formats of score reports that make them more or less useful to educators for identifying students' strengths and weakness and for designing, monitoring, and adjusting instructional programs. The project included a review of assessment reporting practices and research. Field-based educators who are deeply involved in curriculum, instruction, and assessment activities in their schools and school districts participated in two focus groups to provide their insights and suggestions about the substance and formats of various score reporting approaches.

The critical information and features of score reports that might make them especially useful were identified in the first focus group. This information was used to design six score reporting strategies and formats that were reviewed and evaluated by the second focus group. A slightly more detailed description of the project activities is contained in the description of the various sections of this report below, and complete details about the project work are contained in Sections 3 and 4 of the report.

C. Organization of the Report

The report is organized into four sections that follow this introductory section. Titles and brief annotations at the beginning of each section are provided as an advance organizer and guide to the report.

Section 2. Measurement Background, Context, and Previous Research

This section provides a review of the basic concepts and procedures used in score reporting, with examples of commonly used formats and approaches. Five general item mapping strategies are then described. The final subsection presents a review of research on test-score reporting and interpretation.

Section 3. Item Mapping/Reporting Strategy Development Process

The project proceeded through a series of phases that are described in this section of the report. Throughout the project, information and materials related to score reporting and interpretation were collected on an ongoing basis.

The first phase in the process dealt with developing a description of the critical information and features that educators thought would make various score reports more useful and informative. This was accomplished in Focus Group 1, and both the procedures and the results of this first focus group are described in this section of the report.

In the second phase of the process, the results from Focus Group 1 were used to the design and develop six reporting strategies and formats that reflected the information and features identified in the focus group and the review of current practices and research. The procedures and results of this phase are described in this section of the report.

The third phase entailed a review and evaluation of the six reporting strategies and formats. This was accomplished in Focus Group 2. The procedures of both the qualitative and quantitative approaches used with this focus group are described in Section 3. The results of the focus-group review are presented in Section 4.

Section 4. Review and Evaluation of the Score Reporting Strategies and Formats

The results of the review and evaluation of the six reporting strategies and formats are presented in this section of the report. The qualitative review of each of the reporting strategies and formats is presented first. This includes a detailed description of each reporting approach followed by an analysis of the focus group results. Then, the results of the qualitative evaluation ratings of the six approaches are presented and summarized.

Section 5. Summary, Discussion, and Recommendations

The final section of the report reflects on the review of research and the results of this study. The major trends and findings are summarized, key issues are identified and discussed, and final recommendation for developing useful score reports are provided.

Section 2
Measurement Background, Context, and Previous Research

A. Introduction and Section Overview

The interpretation of test results is a complex activity that requires an understanding of a wide range of theoretical and practical strategies and procedures for test development, analysis, reporting, and interpretation. This project focuses on score reporting and interpretation and assumes some general familiarity with basic measurement issues and practices. Nevertheless, a review of certain score reporting and score interpretation issues and procedures seems essential for understanding the work presented in this report.

This section contains:

- Characteristics and Examples of Basic Score Reports.
- Item Mapping and Test Reporting Approaches Based on Item Response Theory (IRT) Scaling.
- Research on Test Score Reporting and Interpretation.
- Discussion and Conclusions.

B. Characteristics and Examples of Basic Score Reports

Score reports can be examined from a wide range of perspectives, and numerous features of score reports can be considered. A framework of eight key characteristics of score reports was developed in order to have a common understanding of basic terms and concepts that will be used throughout this study. This framework was developed from considerable experience with local and state testing programs, a review of numerous score reports from state and commercial assessment programs, and the review of examples from various studies (to be mentioned later in this section). The basic characteristics of score reports that need to be considered in designing a reporting system are shown on Table 1 on the next page.

Table 1

Score Reporting Framework with Features, Options, and Notes

Reporting Feature	Options and Notes
Audience for the Report	Student, teacher, parent, school district, and state Reports are prepared for various audiences and what is contained in the report and how the information is presented may and generally does vary depending on the audience and users of the report.
Scale or Metric for Reporting	Raw score, percentage correct, scale scores, stanines, grade equivalent, and normal curve equivalent The scale or scales in which scores are reported can add clarity or confusion to the score report. It is often simpler to report raw scores or percent correct scores, but these scales do not provide comparability across strands on a single test or between two different tests.
Reference for Interpretation	Norm-referenced, standards-referenced (achievement levels), or both Test results can be interpreted in reference to some normative information, such as percentiles or by reporting how students in the school, district, or state perform on the test. In most states, test scores are reported in terms of content and/or performance standards. Reporting students' test scores in terms of performance achievement levels is proving to be a useful approach.
Assessment Unit	Item, strand (e.g., subscale or subdomain), total test Educators' interest in more instructionally useful information often leads to the request for information about how students perform on individual items or on subsets of items, such as content strands. Strand-level information is commonly reported but has some technical limitations that will be examined later in this report.
Reporting Unit	Student, teacher, school, district, state, nation Score reports are routinely provided for individual students and for different aggregations of students from classroom to the entire nation. Certain features of all reports are the same but each level of report does require different information and approaches.

Error of Measurement	For each unit, metric, and test level combination The precision with which test scores are measured is often reported for performance at the total test level. However, precision as operationalized in terms of the standard error of measurement is not reported as often when strand-level achievement is reported.
Mode of Presentation	Numeric, graphic, narrative Test results can be presented numerically, graphically, or in descriptive narrative form. The best approach for different audiences is not clear, and the use of multiple modes of presentation with some built-in redundancy is often seen in score reports.
Reporting Medium	Print, website-based (static), website-based (interactive) Test results have been traditionally presented in printed hard copy form. This practice will likely continue for some time, but electronic versions supplied via the internet or on CDs are increasingly common.

Most of the features and various combinations of these features are familiar to educators and others.

The basic information that might be contained on a student report is displayed in Table 2, with a hypothetical example of a 34-item test comprised of four strands or subscales.

Table 2

Example of Basic Information in a Student Report

Strand	Number of Items	Number Correct	Percent Correct
Number Relationships	11	10	.91
Geometry	6	4	.66
Algebra	7	4	.57
Measurement	10	4	.40
Total	34	22	.65

The information in Table 2, while commonly reported, is not especially meaningful. The relative difficulty of the subscales is not clear since the raw score metric does not provide a scale in which the strand difficulties have been equated. Scale scores at the strand level equated across all strands would provide a reporting metric on which students' performance could be compared. Because there is no indication of the error of measurement in this table, the extent to which differential performance on the strands exceeds random variation

cannot be determined. In addition, there are no normative or achievement-level connections that could be used to interpret these scores.

These results can be shown graphically as in the bar chart in Figure 1.

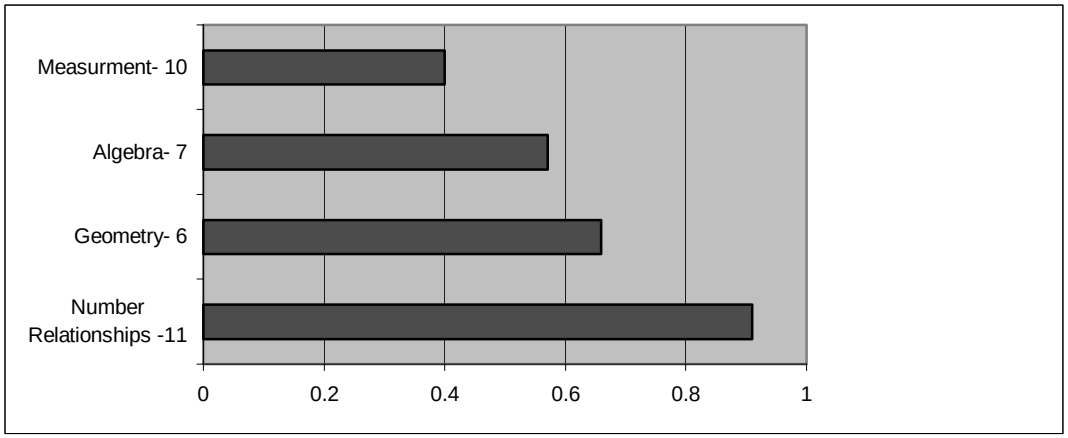

Figure 1. Bar chart illustration of subscale reporting format.

The actual information in the bar chart is not an improvement over the information in the table format; however, if the strand results were presented in equated scale scores, the relative strengths and weaknesses would be apparent.

A student's performance relative to achievement-level categories is shown in Table 3. Such a display assumes some form of comparable scale across the strands. The usefulness of the information in Table 3 would be increased if other information about where in the achievement level a student is located and the errors of measurement or classification consistency were provided.

Table 3

Illustration of Interpreting Subscale Performance at Cut Score-- Shaded Area Shows Students' Achievement Level

Subscale	Basic	Proficient	Advanced
Number Relationships			
Geometry			
Algebra			
Measurement			

An interesting variation on the basic raw score table (Table 2) and the achievement level referenced table (Table 3) is shown in Table 4. This type of table is used by one state to enhance the interpretation of students' strand-level performance. In this table, raw scores at

the strand level are referenced to performance that is expected of students at the Proficient cut score set on the total test.

Table 4

Illustration of Interpreting Subscale Performance at Cut Score

Subscale	# Possible	Weakness	Band	Strength
Number Relationships	11		(5-6)	10
Geometry	6		(2-3)	4
Algebra	7		(2-3)	4
Measurement	10	4	(5-6)	

The cut score for the Proficient level on the total test is interpolated to the strand level using procedures to be described shortly. The strand-interpolated cut score is truncated, and one score point is added to form a band designed to show where students at the Proficient level overall would be expected to perform on the strand. A confidence interval could also be formed by adding and subtracting a standard error from the interpolated cut score.

Students are reported as having a “Weakness” or “Strength” on the strand depending on whether they have scored above or below the Proficient level band, and their raw score on each subscale is reported. This approach has the added meaning of referencing an achievement level, and any achievement level or several could be used.

The lack of comparability in difficulty across subscales is compensated for in the interpolation process and is revealed by the fact that the Proficient level band can have different values, even for subscales with the same number of items. This report, by itself, does not give an indication of the errors of measurement for the individual subscales or the differences between the subscales.

Many score reports attempt to reflect students’ performance at the strand level in terms of achievement levels defined on the total test score and also show the subscale errors of measurement. A basic version of this format is shown in Figure 2 on the next page.

Subscale	Basic	Proficient	Advanced
	Scale Score 200 ---------- 500		
Number Relationships			--X--
Geometry		-----X-----	
Algebra		------X------	
Measurement	---X---		

Figure 2. Subscale performance referencing achievement levels and errors of measurement.

This figure shows a student's performance at the strand level in terms of the scale score metric and in terms of the overall achievement levels of Basic, Proficient, and Advanced. The 'X' on each line represents the students' strand score and the dashes to the left and right of the X represent the 95% confidence interval around the score. The slightly wider intervals around Geometry and Algebra indicate larger errors of measurement, which one might expect because these subscales are often shorter than Number Relationships and Measurement.

These features of score reports for individual students can be carried through to reports for groups of students such as classrooms, schools, and school districts. Many of the same features would be incorporated into group reports with the reporting value generally being a group mean, group standard error, or a group percentage when classification categories are employed.

C. Item Mapping and Test Reporting Strategies Based on Item Response Theory (IRT) Scaling

Operating behind the scenes of most score reports is the set of psychometric procedures known as Item Response Theory (IRT). While there are different IRT models and approaches, they all have in common the capacity to place the performance of the students and the items that students answer on the same scale. Locating people and items on the same scale greatly enhances score interpretation. The IRT approaches enable one to examine the performance of a student and describe the items the student is more or less likely to answer correctly.

In the IRT approach, items that students have a low probability of answering correctly can be described as assessing content the student has not yet learned. Items that students have a high probability of answering correctly can be described as assessing content students have learned at some level of proficiency. Finally, items on which students have some mid-range probability, such as .4 to .6, of answering correctly can be described as assessing content the students are in the process of learning. The probability levels used,

the descriptions of the levels, and strategies of describing the content at various levels will be examined below.

Five item mapping and test reporting strategies that are based on the use of IRT analyses will be described. These include:

- Interpreting a scale by mapping items and variables.
- Interpreting performance at a score point or cut score.
- Interpreting performance for an achievement level.
- Interpreting and reporting scale and achievement levels graphically.
- Interpreting by mapping performance from one set of items to another.

It is useful to note that the terms *item map*, *item mapping*, and *variable map* are all used in the measurement community and are often used with different meanings. In addition, the concept of "mapping " in the IRT approach can also refer to mapping ability estimates from one set of items to another. These distinctions will be clear in the explication of the five approaches.

Interpreting a Scale by Mapping Items and Variables

There is a long tradition of item mapping or variable mapping for measurement practitioners and researchers who work extensively with the one-parameter IRT model. This model is restricted in that test items are represented by only one parameter, namely the item difficulty. Item discrimination and the pseudo-guessing (c_g) parameter are not employed, as they are in the more general two- and three-parameter models. Thus, with this restricted one-parameter model the items can be positioned along a continuum based on their item difficulty and no additional information is needed (or available, save the standard errors of the estimated difficulty). This approach is shown with a hypothetical example in Figure 3 on the next page.

Figure 3 illustrates the variable (item) mapping approach using a five-item basic arithmetic test. The measurement scale is ordered from difficult or hard at the top to easy at the bottom, based on their IRT estimated difficulties. The (hypothetical) abilities of students with raw scores of 1, 2, 3, and 4 and the items with brief content descriptions are shown on the same scale.

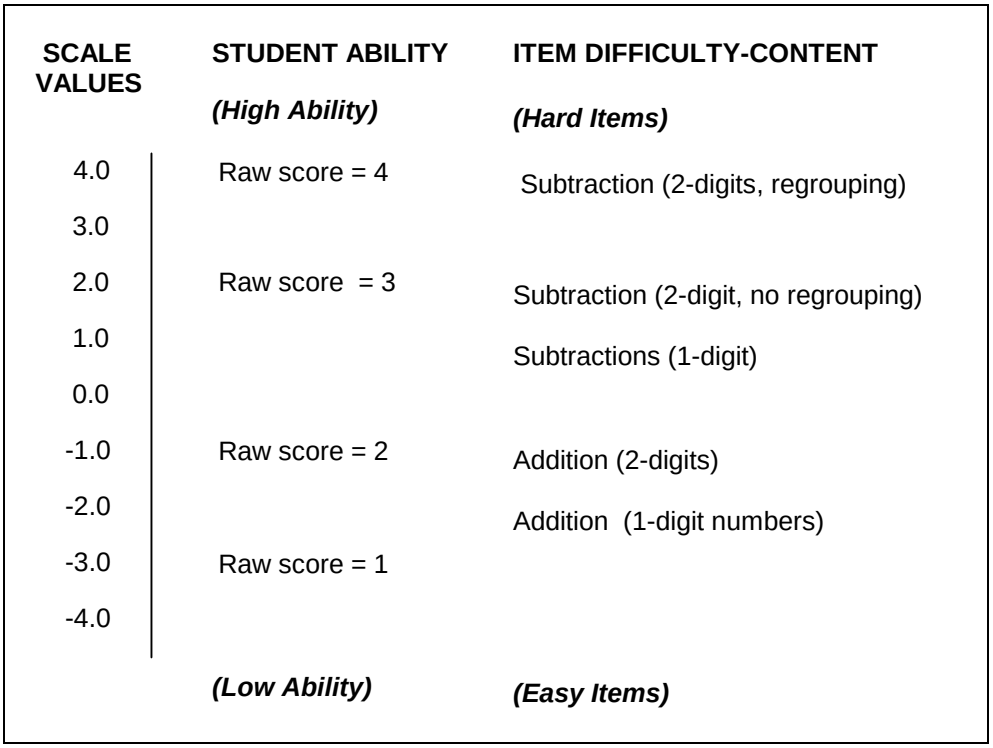

Figure 3. Illustration of variable map based on a five-item arithmetic test.

This example shows several features of "item mapping" or "variable mapping" as it is often described in the context of the one-parameter model for score interpretation. First, the scale can be described in terms of the content of the items that is "mapped" by ordering the items based on their difficulties. Thus, this scale goes from 1-digit and 2-digit addition, to 1- and 2-digit subtraction without regrouping, to 2-digit subtraction with regrouping. The substantive meaning of the scale is manifested in the content of the items.

Second, people can be ordered on the same scale as the items since the raw score on the test is sufficient information for estimating a person's ability and the IRT scale is the same for person ability and item difficulty. Third, the probability that a person with a given raw score/ability will correctly answer an item can be seen on the map. If the raw score, which is the basis for estimating a student's IRT ability, is at the same scale location as an item, then the person has a probability of .5 of responding correctly to that item. As ability exceeds the item difficulty (higher ability students taking relatively easier items), the probability of a correct response increases. Conversely, as the ability becomes less than the item difficulty, the probability of a correct response decreases.

This approach to item or variable mapping has been a standard part of the one-parameter IRT applications as described by Wright and Stone (1979), Wright and Masters (1982), and as seen in such commonly used software as *Winsteps (*Linacre, 1999). Variable map construction and interpretation using this approach has been used extensively in applications of the rating scale and partial credit one-parameter models (Coster, Ludlow,

and Mancini, 1999; Masters, 1982; and Wright and Masters, 1982). The *Journal of Outcome Measurement* and the *Journal of Applied Measurement* focus specifically on this approach to score and scale interpretation and Stone, Wright, and Stenner (1999) in their article, "Mapping Variables," provide a detailed discussion of this approach.

The approach described here, in conjunction with the restricted one-parameter IRT model, can be used with the more general two- and three-parameter models with a small variation. In these approaches, item discrimination and the c_g or pseudo-guessing parameter are taken into account. The picture shown in Figure 3 cannot be drawn directly because the item difficulty is not a sufficient statistic for determining the probability that a student with a known ability will correctly answer an item.

With the more general IRT models, the vertical axis represents the ability scale in IRT logits or a derived scale score and items are then plotted based on a decision about what response probability will be represented in the figure. The response probability (RP) refers to the probability of a correct response for students at a given ability (or the derived scale score that might be shown on a graph or figure). If a RP of .67 is employed, then items are located at the scale score position such that students with the ability at that point have a .67 probability of responding correctly. In the one-parameter case, the RP of .50 is generally used as the default but a RP of .67 or higher could be used.

Interpreting Performance at a Score Point or Cut Score

A second IRT based approach to interpreting students' test scores is mentioned in the introduction and can be seen in the preceding example. Students' scores are the basis for estimating their abilities on the IRT scale. Once the ability of a student or students at a particular score point are known, that performance can be described in terms of what items the students will most probably answer incorrectly or correctly. The content of items on which students have a high probability of answering correctly can be described as content students have mastered; the content of items on which students have a low probability of answering correctly can be described as content students have not yet mastered; and, the content of items on which students have a probability of responding correctly in a middle range, say for example a probability between .40 to .60, can be described as content on which students have some knowledge but have not yet mastered.

Which response probabilities should be used for these interpretations and in other applications, which items at different difficulty levels should be considered, and how the content is best described will be examined in more detail shortly. The key issue here is that this approach allows for at least a probabilistic answer to the question, "What does a score tell us about what a student knows and can do?" In Figure 3, for example, a score of 3 means the student can probably do one- and two-digit addition, has some ability to do one- and two-digit subtraction without regrouping, and probably is not yet able to do two-digit subtraction with regrouping.

Interpreting Performance for Achievement Level Intervals

One of the most useful and popular score interpretation strategies involves developing what are called Performance Level Descriptions (or originally, in NAEP, Achievement Level Standards). This approach starts by setting several cut scores that define various levels of achievement such as Below Basic, Basic, Proficient, and Advanced. Test items or items from a test bank that fall into these achievement levels are then identified if their IRT difficulties fall into these ranges or if they are mapped into the range by their response probability. Content experts identify model or exemplar items for each level and then provide narrative description of the content and task demands that are typical at each achievement level. The basic framework for this approach is shown in Figure 4.

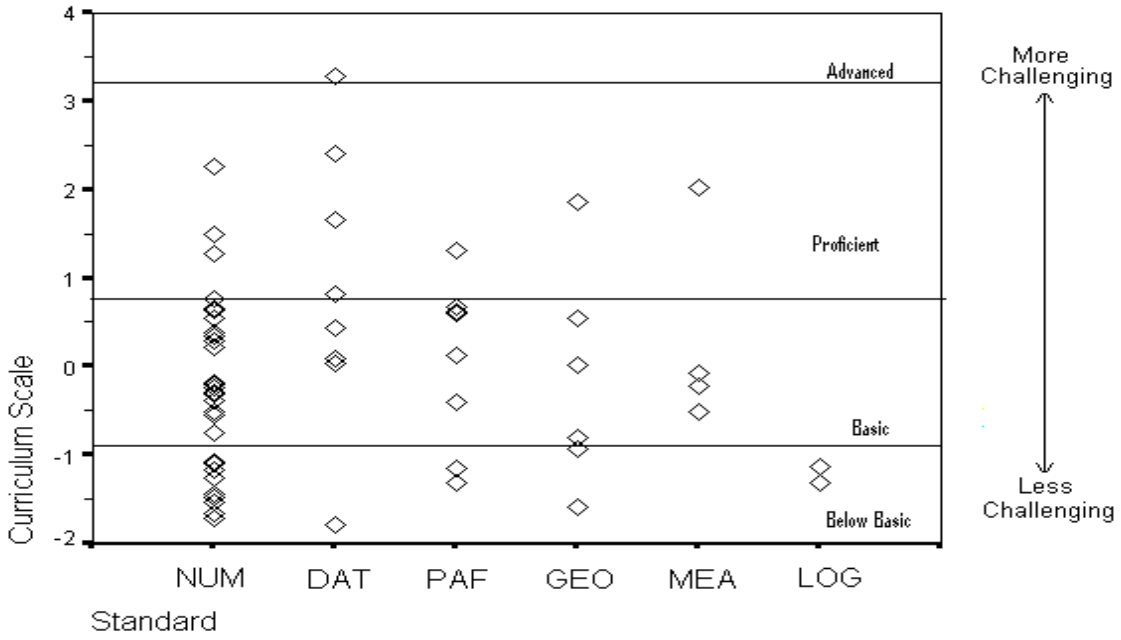

Figure 4. Curriculum map for Grade 5, Mathematics showing achievement levels and content strands (standards).

This example was developed for an application at a large school district in the Southwest and is based on a fifth grade mathematics test taken by approximately 3,000 students. All analyses were done with the one-parameter IRT model. The four achievement levels used were Below Basic, Basic, Proficient, and Advanced. The three cut scores defining the four levels were determined using the contrasting group method (Cizek, 1996; Livingston and Zieky, 1982). The district chose to call the vertical scale the "Curriculum Scale," denoting the fact that items that reflect the district's curriculum could be ordered on the scale. The horizontal axis is used to indicate the content standards: NUM = Numeration; DAT = Data Analysis; PAF = Probability and Functions; GEO = Geometry; MEA = Measurement; LOG = Logic. Each diamond in the figure represents an item from the test. In most settings, the plot symbol would be a curriculum code or a short description of what the item measures.

In the above example, achievement-level descriptions would be developed by examining the items that fall into each level. It is useful to note that the richness of these descriptions depends on the density of items within a strand and within any achievement level. The data in this example show that very little can be said about what students in the Advanced level know and can do that would differentiate them from students in the Proficient level.

The use of achievement-level categories has become a standard feature in virtually all large-scale assessment programs. The early use of the achievement levels of Basic, Proficient, and Advanced in the National Assessment of Educational Progress (NAEP) certainly set the stage for and encouraged the use of such categories. The use of these achievement levels is designed to add substantive meaning to students' classification and to test scores.

Issues in Developing and Using Achievement Levels

A variety of issues arise in the use of achievement-level categories and descriptions. The first issue is to decide how many categories will be defined. Many programs originally focused on three categories. Currently, state assessment programs commonly use four categories: Below Basic, Basic, Proficient, and Advanced; and now some states are considering five categories, which would require four cut scores or performance standards. The number of categories used is partly a policy decision. The decision also needs to consider the test length and, given the number of items on a test form, how many categories can be defined in a way that performance in the categories are actually significantly different from each other.

The second issue to consider is how to actually set the performance standards. Any number of the procedures described in Cizek (2001) for setting standards seems to work adequately for most purposes. The Angoff (1971) procedure and modifications of the Angoff approach have been widely used. More recently, a variety of item mapping procedures such as the *Bookmark* approach (Mitzel, Lewis, Patz, and Green, 2001) have been applied in many state testing programs. The item mapping procedures have in common the use of items ordered based on their IRT difficulties or response probabilities.

The third task that must be addressed is the critical task of developing achievement-level descriptions that define what students in each category know and can do. In the example above, three cut scores are used and these define four achievement levels. An excellent review, analysis, and investigation of the steps involved in developing the achievement-level descriptions is provided by Zwick, Senturk, Wang, and Loomis (2001).

The process begins by identifying items, anchored to the respective cut scores, which are considered to be characteristic of the respective achievement levels. Items that should be considered representative of an achievement level are items that students in that level will probably answer correctly. As described earlier, the probability level used is called the *Response Probability* (RP). But what level of probability should be used to identify the

items? Many researchers and practitioners have used an RP of .50. The logic of the .50 RP is that it marks the point at which 50% of the students will correctly answer the item and, therefore, the item may be considered an exemplar of the achievement level. Others have suggested higher RPs of .67 or .74. Huynh (1998) recommends the use of RPs of .67 or .74, depending on the type of item and IRT model being used. Zwick, Senturk, Wang, and Loomis (2001) also support the use of RP of between .65 and .74, both on empirical grounds and based on the judgment of subject matter experts.

Exactly where the items that define an achievement level should be anchored must also be considered. In many cases, the items are anchored to the ability at the cut score that defines the beginning of the achievement level. Values throughout the range of the interval also can be used and Zwick, Senturk, Wang, and Loomis (2001) report that anchoring to the midpoint of the interval works effectively.

Critical to the work described in this report is a consideration of the intended purpose to which the achievement-level descriptions will be applied. The procedures developed for achievement-level descriptions have been designed primarily for accountability purposes and are used to report what students know and can do. A different purpose, however, might have a more diagnostic intent and be more instructionally related. With an instructionally related purpose, achievement-level descriptions might be used to report what content objectives students should currently be taught and what objectives might be taught next. For this instructionally related purpose, RPs in the range of .4 to .6 might be more appropriate than RP's of .67 or .74. The higher RP's are used to ensure that what is described reflects content that students have mastered and that instruction on this material would then seem unnecessary.

Interpreting and Reporting Scales and Achievement Levels Graphically

The information contained in Figures 1 and 2 has a value beyond facilitating the explanation of various content- and achievement-level interpretations. These figures, or figures like them, can also be used as the mechanisms for presenting and interpreting test results. The review of research on score reports (presented in subsection D2) shows that many practitioners find the use of graphical displays helpful in interpreting test results. The graphical displays in most score reports, however, are fairly conventional and are used to convey such basic information as number or percent of items correct, scale scores, or scale means. As will be seen, some score reports are beginning to include graphical representation of students' performance in terms of achievement-level reporting (e.g., Table 3 and Figure 2). However, score reports showing the variable map as in Figure 1 or the achievement levels by strands as in Figure 2 have not been explored as formats for reporting test results.

Interpreting Performance by Mapping from One Set of Items to Another

A valuable tool in the interpretation of students' test scores involves using test scores to estimate students' expected or likely performance on items or sets of items that they have

not yet taken. This can be used to estimate performance from a test to the likely performance on some proposed tests or on an entire bank of items that might be of interest. The same approach can be used to estimate how a student at a particular score point can be expected to do on individual items or a subset of items.

The most common application of this general strategy involves identifying the IRT ability required to attain some cut score or performance level (e.g., Proficient) set on the total test and using this ability estimate, with the strand-level item parameters, to estimate the equivalent cut score on the content strand as in Table 4. This approach allows for the interpretation of achievement-level standards that have been defined on the basis of the total test in terms of the equivalent levels on test subscale or strand. Practitioners use this procedure to superimpose achievement-level standards from a total test to content strands; e.g., geometry, measurement, and algebra in mathematics or communication, reading and writing in language arts.

D. Research on Test Score Reporting and Interpretation

Previous research on score reporting and test interpretation is not extensive as compared with research on other topics that measurement and psychometric communities have explored. Such reported research, however, seems to present a fairly consistent although somewhat bleak picture of the effectiveness of score reports to communicate meaningful information to various stakeholder groups.

The review of the research related to score reports for this project is developed from information at the following four levels.

- A state-level case study
- A recent multi-state review of score reports
- The national review panel
- Research review summary and trends

As the reader will see, there is considerable consistency in the findings concerning score reports and interpretation across these four perspectives.

D1. A State Level Case Study

A very useful four-stage process for reviewing and redesigning the reporting system for the Connecticut Mastery Test (CMT) that could be applied in any number of settings is described by Forte Fast and Tucker (2001). Stage one involved a review of existing state assessment reports and state and federal reporting requirements. Stage two extended the study to a review of assessment reports from other states. Stage three involved a series of focus groups held around the state that included parents, teachers, and administrators. In stage four, information from all sources was used to redesign the student, classroom, school, district, and state reports.

Some of the findings of the Connecticut study might be idiosyncratic to Connecticut. However, many of the findings address general issues that apply directly to other states and the observations and suggestions reported in this study are useful as illustrations of the types of issues that should be considered when designing test reports.

Forte Fast and Tucker (2001) reported the comments of different groups who were asked to respond to different score reports. This approach is, in itself, an important model because it shows that the expectations and needs of different groups must be considered when examining the effectiveness of different score reports. Comments of the different groups in discussing the score reports as reported by this study are shown below.

Parent and Teacher Review of Individual Student Reports

- *Use larger font.*
- *Personalize the report by using the student's name.*
- *Include comparative information to help me interpret my child's score. For example, how did my child perform compared to others in his school or others in the state?*
- *Include graphical representations of data to help clarify meaning.*
- *Make clear what additional information is available and how it can be obtained.*
- *If terms are used that may be unfamiliar to a reader who does not work in education (e.g., holistic), provide an explanation of those terms.*

Teacher and Administrator Review of Classroom Level Summaries and Diagnostic Reports

- *Don't use the slanting format for student names. It is hard to read and difficult to line up with other reports.*
- *Boldface type is not enough to indicate which objectives a student has mastered.*
- *Create a … class report with students' names placed in cells according to their performance level on the test.*
- *Reports should flag areas that need additional work.*
- *Don't focus so much on the "goal." We need information about all score bands.*

School and District Administrator Review of School, District, and Statewide Reports

- *Use colors or patterns with a high level of contrast so that they copy well in black and white.*
- *Provide more graphics to help us present our data. Otherwise, we have to produce our own.*
- *We need a better way of re-analyzing our data. MTIS [the state electronic data/report delivery system] needs to be more user-friendly and needs better graphics.*
- *The percent above the remedial standard is important to us.*

- *Stacked-bar graphs showing the percentage of our students at each performance level are very helpful.*
- *Include both graphics and numbers. Both are important.*

These comments, offered by different stakeholder groups examining reports at different levels, provide ideas and suggestions for designing test reporting documents and systems. The results of this study, looking across the various levels of reporting, reflect the need for assessment program designers to consider at least the following broad categories of assessment report features:

- Format features (e.g., type face, font size, bold, use of color, general layout)
- Graphical displays
- Numeric displays
- Normative information
- Detail and specificity (especially about strengths and weaknesses)
- Support materials
- Glossary of terms
- Directions to supplementary information
- Easily reproduced materials

The Connecticut four-stage process used to collect the information in this study may also be useful for revising and refining reporting systems in other settings. The use of focus-group meetings held with various stakeholder constituencies to review score reports for students, schools, districts, and the state seems especially valuable.

The direction of this work is continued in a publication by Forte Fast, Blank, Potts, and Williams (2002) designed to help states and local agencies meet the reporting requirements of NCLB. This publication contains guidelines and examples that state and local agencies can use to improve the effectiveness of their score reporting procedures and formats.

D2. A Multi-State Review of Score Reports

Goodman and Hambleton (2003) provide a valuable resource for researchers and practitioners interested in the study and improvement of score reporting. In addition to examining score reports, their work examines assessment interpretative guides and materials but that aspect of the report is not reviewed here.

The study examines the score reports from 11 states including Connecticut, Delaware, Louisiana, Massachusetts, Minnesota, Missouri, New Jersey, Pennsylvania, Virginia, Wisconsin, and Wyoming. Also included in the study are score reporting materials from Harcourt Educational Measurement (Stanford-10), CTB/McGraw Hill (Terra Nova, 2nd Edition) and Riverside Publishing (Iowa Test of Educational Development). The score

reporting materials from the Canadian province-wide assessments of British Columbia and Ontario were also reviewed in this study.

The study focused on score reports for individual students and began with an iterative content analysis to review, analyze, and summarize each of the student reports. A category coding system was created to address the key features of the score reports taken as a group.

The synopsis of the major findings of the Goodman and Hambleton study follow the general outline of their report. The information presented below is taken directly from their report with minor rephrasing and reorganization.

Features that Make Score Reports More Readable

The review of the score reports show that certain features of the reports seem to make them more readable. These include:

- Using headings and other devices, such as boxes, lines, white space, and perhaps color to organize reports.
- Using a highlight section that provides readers with an overall summary of results.
- Using graphical displays to draw readers' attention to major results, showing how students performed overall or on major components of the test.
- Designing reports for specific audiences to meet the different needs of different groups.

Personalized Score Reports

Several score reports examined by Goodman and Hambleton employed a strategy to personalize the reports by using the student's first name in several places in the report. This seems to be a useful score report feature, but it does require an accurate name file that can be sorted and matched with a report data file. The results should look like more than just a name dropped into a fixed space by accommodating names of different lengths.

Features that Appear to Add Meaning for Intended Users of Student Score Reports

The major point of studying score reports is to develop reports that will be understandable to those who read and use them (as required by NCLB). The score reports reviewed show a number of design features that make reports more meaningful to students, parents, and teachers. Goodman and Hambleton report the importance of:

- Describing the skills and knowledge assessed by the test.
- Describing the expected levels of performance on the test through well-defined performance levels.

- Describing the skills and knowledge a student possesses or does not yet possess through use of performance levels or diagnostic information such as subdomain results and descriptions of specific strengths or weaknesses of particular students.
- Reporting the results of relevant comparison groups (e.g., other students in the school, district, and state).
- Reporting results in multiple ways (e.g., using numbers, graphics, and narrative text).

Reporting Results in Relation to Performance Levels

Following the approach that evolved with the NAEP program, state assessment programs generally have performance standards that define various achievement levels. The findings of Goodman and Hambleton show that there are several important features of score reports that present the results of students' performance in relation to the performance levels. These include:

- Providing a general description of the performance levels.
- Displaying results graphically with accompanying text .
- Presenting some form of normative information, such as the percent of students at different performance levels for a school, district, or the state.
- Showing or reporting how close a student is in achieving different performance levels.
- Providing information about errors of measurement when reporting students' performance levels.

All of these features are seen as providing more meaningful information about students' performance in relation to various performance levels. The concern about information regarding precision or standard errors of measurement is of particular interest in the current study. There are several aspects of the performance-level reporting approach in which precision is relevant. Certainly, knowing the standard errors of measurement for the measurement instruments and for any subset of items such as a subdomain or strand is critical. In addition, precision in the classification of students might also be reported. The classification consistency for a single test administration can be estimated using the procedures of Huynh (1978, 1979) as will be illustrated in Section 4, Part B3 of this report. Finally, if students are assigned to achievement levels for different content strands for the purposes of identifying strengths and weaknesses, then the reliability or precision of the differences in strand-level performance should be investigated.

Reporting Diagnostic Information – Subdomain (Strand) Scores

State assessment programs are generally established for two purposes: accountability and instructional support. The instructional support function suggests that the programs will provide diagnostically useful information to teachers and other educators that can be used to review and revise school programs. The diagnostic information from state assessments takes the form of subdomain or strand-level information. In mathematics, for example,

common subdomains or strands include Numbers and Operations, Geometry, Algebra, Measurement, Patterns and Functions, and Data Analysis and Probability.

Goodman and Hambleton found that most large-scale assessment programs provide strand-level information in the form of raw scores, percent correct scores, or percentile scores. Another common metric for reporting strand-level performance is the use of scale scores. Additional meaning can be added to students' strand-level information by reporting some form of comparative information, such as district or state performance on the strands.

Providing information about the precision of the measurement at the strand level is an often neglected but critical feature essential in interpreting the results. The reliability of strand-level performance will be examined in detail in this study in Section 4, Part C. Reporting performance on the strands is often the basis for developing a profile of strengths and weaknesses for a student; e.g., the student is stronger in certain strands and weaker in others. Such interpretations invite inferences about differences in strand-level performance and in such cases, the precision of the differences should become a matter of concern.

Weaknesses

In their summary, Goodman and Hambleton (pp. 55-56) observe that while many features of the score reports they studied seem useful and others are promising, certain weaknesses or potential weaknesses were noted. These include the following:

- Excessive amounts of information (e.g., multiple types of comparable scores) were included in some reports, and essential pieces of information (e.g., the purpose of the test, information about how the results will be and should be used) were not provided in others.
- In many instances, information regarding the precision of test scores is not provided, making the results appear more accurate than they are.
- While not widespread, statistical jargon such as standard errors, NCE scores, and Lexile scores were present in more than a few reports.
- Key terms, including the critical performance levels, were not always defined in the reports or interpretive guides, leaving the interpretations up to users, many of whom would be quite unaware of the proper interpretations to be made.
- Efforts to report a large amount of information in a small physical space resulted in reports and interpretive guides that appeared dense and cluttered. Small font size was a common cause of concern across many reports and guides.

General Recommendations for Score Reporting

Based on their review, Goodman and Hambleton offer the following recommendations for designing score reports.

- Score reports should be clear, concise, and visually attractive.
- Score reports should include easy-to-read text that supports and improves the interpretation of charts and tables.

- Care should be taken to not try to do too much with a data display (i.e., displays should be designed to satisfy a small number of pre-established purposes).
- Devices such as boxes and graphics should be used to highlight main findings.
- Data should be grouped in meaningful ways.
- Small font, footnotes, and statistical jargon should be avoided.
- Key terms should be defined, preferably within a glossary.
- Reports should be piloted with members of the intended audience.
- Consideration should be given to the creation of specially designed reports that cater to the particular needs of different users.

D3. National Review Panel

The National Education Goals Panel (NEGP,1998) provided suggestions about how states could more effectively communicate with parents about state standards and state assessment and how state score reports could be enhanced (Goodman & Hambleton, 2003). The NEGP report provides a number of useful suggestions about how schools could work more effectively with parents. Of special interest for the current project is the focus group that was used in the NEGP research. A focus group comprising 11 parents from across the United States was used to gather information about what parents liked and disliked about various score reports. In the focus group, parents were asked to review and comment on six individual student reports produced by commercial test publishers.

In general, parents involved in the study:

- Appreciated explanations of what the scores on the test meant.
- Liked to be able to tell at a glance how their child performed.
- Liked to see subtest scores and descriptions of the skills assessed by the test.
- Appreciated learning what could be done to improve a student's score.

Parents did not like reports that:

- Were too technical (e.g., containing statistical jargon and complex definitions).
- Did not give recommendations on what they should do with the test results.
- Used small fonts that made parts of the reports difficult to read.

This study provided results that are informative and consistent with other research on score reporting. Of particular importance is the use of a parent focus group as a basic data-collection procedure. It seems increasingly clear that there is considerable value in asking stakeholder groups to review drafts and prototypes of score reports that will eventually be used to supply them with test result information.

D4. Research Review Summary and Trends

An excellent and very current review of literature related to score reporting is provided by Goodman and Hambleton (2003). The major finding of this review is that *many users of assessment data have difficulty interpreting and understanding results presented in large-scale assessment reports* [italics added]. This general conclusion is based on Hambleton (2002); Hambleton and Slater (1997); Impara, Divine, Bruce, Liverman, and Gay (1991); Jaeger (1998); the National Education Goals Panel (NEGP, 1998); the National Research Council (NRC, 2001); and Wainer, Hambleton, and Meara (1999), as cited in Goodman and Hambleton (2003).

This general finding is based on research reviews that focused substantially on NAEP score reporting approaches and formats. Nevertheless, the results of this work apply to today's issues of providing informative and useful score reports for individual students and other levels of score reporting. Among the problems seen in score reports, as critiqued in the literature, are:

- Reports assumed an inappropriately high level of statistical knowledge.
- Statistical jargon confused and even intimidated some users.
- Technical terms, symbols, and concepts were required to understand the message underlying even simple data.
- Technical symbols were misunderstood or ignored by many users of the reports.
- Too much information made it difficult for readers to find and extract what they really want to know.
- The inclusion of overly dense displays was challenging to those reading the reports.
- Graphical alternatives to textual and tabular formats were not used often enough.
- Increased clutter or perceptual inaccuracies sometimes occurred when displays were redesigned for easy access (e.g., using three-dimensional bar and pie charts).
- Reports lacked descriptive information (e.g., definitions and concrete examples) that would have helped provide meaning to the assessment results.

(*Goodman and Hambleton, pp.8-9*)

Goodman and Hambleton report a set of general principles that they have extracted from recent literature of score reporting that they cite in their report that includes Hambleton (2002); Hambleton & Slater (1997); Jaeger (1998); NRC (2001); Snodgrass & Salzman (2002); (Wainer, 1997a); Wainer et al. (1999); and Ysseldyke & Nelson (2002). The literature relating to the visual display of quantitative information include Tufte (1983, 1990); Tukey (1990); Wainer (1990, 1992, 1997b) and; Wainer & Thissen (1981). These principles include:

- Making the report readable, concise, and visually attractive.
- Keeping the presentation clear, simple, and uncluttered.

- Not trying to do too much with a data display (i.e., displays should be designed to satisfy a small number of pre-established purposes).
- Including text to support and improve the interpretation of charts and tables.
- Minimizing the use of statistical jargon.
- Including a glossary of key terms.
- Using bar charts to facilitate comparisons.
- Grouping data in meaningful ways.
- Using boxes or graphics to highlight main findings.
- Avoiding the use of decimals.
- Using color in a purposeful manner (given the potential for misuse, however, the general use of color was not universally recommended).
- Piloting the reports with members of the intended audience.
- Creating specially designed reports for different audiences.

(*Goodman and Hambleton, pp. 9-10*)

E. Discussion and Conclusions

There is no simple summary of features and formats that make score reports informative and meaningful for various stakeholder groups. The summary of Goodman and Hambleton's results and the research literature summary contain a list of every score report feature that has been identified as effecting score report interpretation. These summaries of score report strengths and weaknesses and the general principles for score report design should be used at the planning and development stage when score reports are being conceptualized, designed, and first drafted.

The review of score reporting literature and practice reveals the use of focus groups to evaluate various score reports designed for different audiences. The use of parent, teacher, and community focus groups to pilot test score reports is recommended as a valuable step in developing informative and useful score reports. It seems reasonable to recommend field testing score reports with their intended audiences. In the field of educational measurement, no one would think of using a test item that had not been thoroughly reviewed and field tested.

Finally, it is important to be clear that this review of score reporting literature and practices did not address the actual use of score reports. The review reported what researchers and practitioners thought and said about various score reports, not how users of the reports interpreted the data or what practitioners actually did with them. A different line of research might involve researchers visiting schools and school district offices to observe and interview students, teachers, district personnel, and parents. The purpose of this line of research would be to describe how the information in various score reports was interpreted and actually used in schools and school districts.

Section 3

Item Mapping/Reporting Strategy Development Process

A. Introduction and Section Overview

The original project plan called for a review of the background research and practice, the development of several prototypes of item maps and score report formats, and then a focus-group discussion in which South Carolina educators would review and comment on the several proposed item maps and reporting approaches. As planning for the project proceeded, the role of the practitioner's focus group was reconsidered.

In the original plan, the educators in the focus group were placed in a reactive role in which they would comment and critique the formats proposed for their review. Their input was to be used to revise the item mapping and score reporting formats as presented to them, and these revised formats would be the basis for the final project report. In this approach, the educators would have no say about which score reporting approaches would be considered but would have been responding to approaches proposed by the research team.

The project work plan was modified to give educators a more direct role in shaping the design of the initial item mapping and score reporting formats. The revised plan was developed in consultation with the South Carolina Office of Assessment. The major change in the work plan involved the use of two focus groups instead of one. The first focus group used an inductive, open-ended approach to provide direction, a framework, and suggestions for designing and developing the proposed item mapping and score reporting approaches. The second focus group used a deductive approach in which the participants were presented with six score reporting formats and were asked to review them and explicate their strengths and weakness.

Part B describes the development of an item mapping/report strategy and format design characteristics and the suggestions from Focus Group 1. Part C presents the procedures for the development and production of item mapping/score report strategies and formats, and Part D describes the process used to review the proposed item mapping and score reporting formats in Focus Group 2.

B. Development of Reporting Strategy/Format Design Characteristics (Focus Group 1)

The initial focus group with South Carolina educators was held in February 2003. The following describes the participants for the meeting, the procedures followed, and the key results.

Participants

The Director of the Office of Assessment, in consultation with staff and others, selected educators for this meeting. Educators were selected who could represent different types of schools, regions of the state, constituencies, and educational perspectives. The fourteen participants included teachers, district and state curriculum coordinators, and district and state research and assessment directors and specialists. The participants and their affiliations are listed in Appendix A.1.

Focus Questions

Participants in this process were asked to consider and discuss the following questions:

- *What information from PACT assessments reported at the district, school, and classroom levels would be most helpful in developing curriculum and planning instruction?*

- *What should PACT assessment reports contain and look like to be most useful at the school and district levels?*

These questions were intentionally left open-ended in order to encourage participants to offer their own ideas rather than asking them to react or respond to ideas or suggestions from the Department of Education or the researcher.

Discussion Procedures

The focus-group participants were given an orientation to the project and the expectations for their participation. They were reminded that they served in an advisory role, that their work was the beginning of a process to explore ways to provide the most useful information from PACT, and that their suggestions would be reviewed by other people and groups framed against a variety of constraints.

Participants were asked to discuss the two focus questions in two subgroups in order to facilitate participation. Subgroups were arranged to be representative of the group as a whole. Participants were invited to discuss the two questions in any order they wished and were encouraged to discuss any other related issues thought to be relevant to the conversation. The subgroup discussions were observed by SCDE staff (some of whom participated) and the researcher, who answered questions.

The conversations in both subgroups were lively and all members of the groups participated actively. Certain common themes emerged quite clearly and quickly in both subgroups. Participants were asked to record the ideas, issues, or suggestions they believed were the most important on a “Committee Member Response Form” supplied for this purpose. After the participants made whatever individual notes they wished, each group selected a facilitator to record the key points on a flipchart. Taking turns, each person offered her or

his top three suggestions or issues. A check was placed next to any point that has already been listed.

Participants reconvened as a group of the whole to review and discuss the major findings that each subgroup recorded on the flipcharts.

Major Findings

The two subgroups recorded a total of 21 summary points on the flipcharts with considerable overlap and nearly total endorsement of all points. Two major substantive themes emerged from a review of the participants' comments. They saw a need for *greater specificity* in reporting students' performance and the need for *more meaningful substantive descriptions* of what scores and achievement levels indicate about what students know and can do. A third category of comments related to more general features of the PACT program. The recorded points from the flipcharts are listed below under the three major summary headings. Certain points are listed under two headings if they seemed to reflect two issues. It is important to remember in reviewing these results that the process of this focus group was open-ended; participants were not given any suggestions about what they might propose in response to the focus questions.

<u>More specificity in reporting students' performance</u>

- *Provide as much specificity at the lowest level of content possible.*
- *Define specific differences between Basic and Proficient.*
- *Provide item analyses to compare school, district and state.*
- *Display student performance in reference to cut points, comparing students, class, school, in the state at the strand level.*
- *Provide descriptions for items, objectives, strands, and standards.*

<u>More meaningful substantive descriptions of scores and achievement levels</u>

- *Provide descriptions of items, objectives, strands, and standards.*
- *Provide strand analysis and/or score interval (achievement level) descriptors.*
- *Describe, based on item analysis, what a typical student at various achievement levels (BB, B, P, and A) can do.*
- *Define specific differences between Basic and Proficient.*
- *Have a group review the results annually and explicate in detail what the results mean in a way that communicates to all.*
- *Create uniformity and clarification across content areas in definition of strands, goals, standards, etc.*
- *Provide information that will improve curriculum and instruction in areas of weakness.*
- *Identify pivotal standards.*

- *Cluster standards together when there are meaningful combinations.*
- *Report statewide weaknesses.*
- *Simplify the language.*

PACT Program

- *Create uniformity and clarification across content areas in definition of strands, goals, and standards, etc.*
- *Disaggregate data by student, school, and grade.*
- *Provide more specific information about the broader assessment system.*
- *Re-roster data and provide two reports in the fall, one for the spring classes and one for the fall classes.*
- *Provide prescriptive information based on standards.*
- *Provide information that will improve curriculum and instruction in areas of weakness.*
- *Simplify the language.*
- *Add a formative assessment component to the system.*
- *Provide more general information about the test, e.g., readability.*
- *Include 45-day enrollment information in reports.*

C. Design and Development of Item Mapping/Score Reporting

Strategies and Formats

The second step in the process involved reviewing the major findings of the focus group and synthesizing these results with guidelines suggested by the measurement literature and by models for item mapping and score reporting from other settings (see Section 2). The Director of the Office of Assessment, psychometric staff, and the researcher had an extensive debriefing session following the focus group. Debriefing/working sessions with the Office of Assessment staff continued through the next day and were followed by extensive phone and e-mail consultations.

The purpose of these working sessions was to design and develop item mapping approaches, procedural strategies, and score reporting formats that were responsive to the needs expressed and the issues raised by the focus-group participants and to ensure that they are psychometrically sound and consistent with recognized measurement practices.

The generation of possible reporting strategies and formats resulted from the review of research and other documentation on reporting results for large-scale assessment programs and from familiarity with state practices for reporting assessment results. The focus group results, and the working sessions that followed the focus group, were also used in the generation of reporting strategies and formats.

The Reporting Strategies and Formats

The review activities led to the consideration of the following six reporting strategies and formats. The descriptions of the strategies and formats provided here is brief because very detailed sections with working examples will be provided in the next section.

1. Item Content Objective Mapping – Graphical mapping of the content objectives associated with each item from a test form, multiple test forms, or the item bank in an ability/item difficulty scale with achievement-level cut scores reported.

2. Achievement Performance Level Narrative – Description of the content objectives assessed by items at the various achievement levels, e.g., Below Basic, Basic, Proficient, Advanced.

3. Strand Achievement Levels for Individual Students – Mapping achievement-level cut scores from the total test level to subscales or strands/areas for individual students and reporting by achievement level.

4. Strand Achievement Levels for Groups – Mapping achievement-level cut scores from the total test level to subscales or strands/areas for groups such as schools or school districts and reporting by achievement level.

5. Observed, Expected, and Differences in Strand and Item Performance for a Group – Observed strand/area and item performance (proportion answering correctly) for schools or districts relative to the proportion expected to answer correctly based on the groups' mean performance on the total test.

6. Observed, Expected, and Differences in Strand and Item Performance at the Achievement Level Cut Scores – Observed strand and item performance (proportions answering correctly) for schools or districts in comparison to the proportion students in the state who are expected to answer correctly at each achievement-level cut score.

Production of Item Mapping and Score Reporting Formats

Staff in the Office of Assessment, working closely with the researcher, developed materials to serve as examples of item maps and scoring reporting formats for the focus group. The materials were developed to reflect, as closely as possible, the designs and guidelines developed after the initial focus group. The development of the item content objective narrative description of the achievement levels (Strategy 2) employed a modification of the NAEP procedure. The details of the procedures will be described in Section 4.

Statewide data from the 2002 third grade operational mathematics assessment were used to produce the materials. The narrative description of the achievement levels was also produced for eighth grade English/Language Arts, using the complete bank of items. The materials were reviewed and edited on site in South Carolina just prior to Focus Group 2 by the researcher and revised by the staff accordingly.

D. Review and Evaluation of the Score Reporting Strategies and Formats (Focus Group 2)

The second focus group session with South Carolina educators was held in March 2003 to review and evaluate the proposed item mapping and score reporting strategies and formats. The following describes the participants, the procedures followed, and the key results.

Two different approaches were used in reviewing and evaluating these score reporting strategies with Focus Group 2. *Qualitative* data were collected to obtain comments and suggestions. In addition, *quantitative* data were collected by having the focus group participants rate the utility of each of the six score reporting formats.

Participants

The Director of the Office of Assessment, in consultation with staff and others, selected the participants for the second focus group. Educators were selected who could represent different types of schools, regions of the state, constituencies, and educational perspectives. The focus group was expanded to 21 participants including teachers, principals, district curriculum and research/assessment directors, and state curriculum and research/assessment specialists. Many of the participants from Focus Group 1 also participated in Focus Group 2. The participants and their affiliations are provided in Appendix A.2.

Review of the Proposed Score Reporting Strategies and Formats

For the qualitative focus group review, a set of explanatory materials was developed for each item mapping and score reporting strategy. These materials contained a description of each strategy that was used by the researcher to explain the strategy to focus group participants. Examples of each report format were also provided to the participants.

Participants were given much the same orientation as was provided to the first focus group. They were then charged with the specific tasks of reviewing the six approaches to score reporting. Participants were asked to review each of the six prototypes for item mapping and score reporting and discuss and answer the following questions for each approach:

- *Will a school or school district find this information helpful?*
- *How could a school or school district use this information?*
- *Could this information be modified to be more informative or useful?*
- *How can this information be best presented?*
- *Might there be any problems in how this information is used?*

The researcher presented each of the six item mapping and score reporting strategies and an example of the report format using the explanatory materials that were developed for the presentation. Participants were invited to ask questions, and these were discussed and answered during the presentation.

Participants reviewed each reporting strategy and format individually and wrote comments and suggestions about the strategy on the comment sheet provided. Then, they discussed each strategy and format in one of three subgroups. Each subgroup selected a facilitator to record the key points on a flipchart. Taking turns, each person offered her or his top three suggestions or issues. A checkmark was placed next to any point that had already been listed. These flipchart results were transcribed and summarized.

The whole group then reconvened to review and discuss the major findings of each subgroup as recorded on the flipcharts and in individuals' notes.

During the subgroup and whole group discussions, participants continued to record their individual comments and suggestions. These individuals' notes were collected, transcribed, and used in the analysis of the qualitative focus-group results.

Section 4 presents the review process, materials, and focus-group comments and suggestions for each of the six item mapping and score reporting strategies and formats.

Evaluation of the Reporting Strategies and Formats

In addition to the qualitative focus group data, quantitative ratings were collected to determine the focus group's evaluation of the usefulness of the six reporting strategies and formats. Each participant rated the strategies at two times in the review process--after the facilitator had described all of the procedures but before any group discussion and sharing took place, and then again at the end of the entire focus group process. The individual rating and comment forms are provided in Appendix D.

Participants applied the rating scale from two perspectives for each item mapping/score reporting strategy. The first ratings were completed from the perspective of a classroom teacher and the second ratings were from the perspective of a district administrator.

As mentioned previously, participants first rated the report strategies and formats during the time they had to review and write comments and suggestions about the six strategies and formats. After the subgroup and whole group discussions of the strategies, each focus group participant rated the six score reporting strategies and formats again. The results and analyses of these evaluation ratings are provided and discussed in Section 4, Part C.

Section 4

Review and Evaluation of the Score Reporting Strategies and Formats

A. Introduction and Section Overview

The focus group procedures described in the preceding section led to the development and review of the following score reporting strategies and formats.

1. Item Content Objective Mapping – Graphical mapping of the content objectives associated with each item from a test form, multiple test forms, or the item bank on an ability/item difficulty scale with achievement-level cut scores reported.

2. Achievement Performance Level Narrative – Description of the content objectives assessed by items at the various achievement levels, e.g., Below Basic, Basic, Proficient, Advanced.

3. Strand Achievement Levels for Individual Students – Mapping achievement-level cut scores from the total test level to subscales or strands/areas for individual students and reporting by achievement level.

4. Strand Achievement Levels for Groups – Mapping achievement-level cut scores from the total test level to subscales or strands/areas for groups such as schools or school districts and reporting by achievement level.

5. Observed, Expected, and Differences in Strand and Item Performance for a Group – Observed strand/area and item performance (proportion answering correctly) for schools or districts relative to the proportion expected to answer correctly based on the groups' mean performance on the total test.

6. Observed, Expected, and Differences in Strand and Item Performance at the Achievement Level Cut Scores – Observed strand and item performance (proportions answering correctly) for schools or districts in comparison to the proportion of students in the state who are expected to answer correctly at each achievement-level cut score.

As mentioned in Section 3, two different approaches were used in reviewing and evaluating the score reporting strategies and formats with Focus Group 2. *Qualitative* data were collected to obtain comments and suggestions from the users of the score reports. In addition, *quantitative* data were collected by having the focus group participants rate the utility of each of the six score reporting formats.

The *qualitative* focus group review results will be presented in Part B. The results for each of the six score reporting formats are organized in the following way. Quantitative results are shown in Part C.

- Introduction – Overview of the score reporting procedure.
- Description and Explanation of the Reporting Strategy and Format – In presenting each reporting strategy, the following information was provided to participants.
  - A description of the procedures that were used to generate the information for the score report.
  - An example of the score reporting format with sample data.
- Focus Group Results – The focus group participants' written comments from their individual response sheets and subgroup discussion recorded on chart paper are presented and categorized as follows:
  - Questions
  - Strengths
  - Weaknesses
  - Suggestions

  The comments of the focus group members are presented as they wrote them. The researcher's comments, added in some cases for clarification, are presented in [brackets]. Subgroup statements are not presented if they are a verbatim reiteration of a person's statement as recorded in an individual's notes.

  After the participant comments about strengths, weaknesses, etc., are presented, a brief summary and discussion about the comments are provided.

- Discussion and Conclusions - A summary and discussion of the review results are presented with conclusions and suggestions for further investigation of the approach and/or reporting format.

Quantitative results and analyses of the focus group's ratings of the six score reporting formats are presented in Part C. These analyses include the mean of the focus group ratings and a rank ordering of the mean ratings of the reporting formats.

B1. Item Content Objective Mapping

Introduction

The graphic mapping of items, sometimes referred to as an *item map* or a *variable map*, was reviewed in Section 2. The strategy presents a rectangular graphical display that shows the difficulty of each item and the ability of students at the various cut scores. An example of an item content objective map is shown on page 38. The vertical axis represents the measurement scale reported in the scale-score metric. The achievement level at each cut score is shown as a horizontal line across the page. These lines divide the scale into four groups, namely Below Basic, Basic, Proficient, and Advanced.

The horizontal axis is used to locate the strands as nominal categorical variables. At each strand location, the items that measure that strand are shown by locating the items' positions on the vertical scale score difficulty dimension and providing a brief (one- or two-word) description.

A number of variations on this map can be considered. Instead of plotting each item each year, the map could be used to plot the mean for each objective over the history of the assessment program. This type of plot would be more stable since the mean of the items for an objective would be plotted. However, such a plot might not be as useful to educators as a current-year item-level map. This mapping procedure could be used to show the entire item bank for a content area and grade. Such a map might be useful in showing the density of the bank's items for the various strands (and possible objectives) relative to the cut scores.

Description and Explanation of the Reporting Strategy and Format

The information on the item content objective mapping strategy that was presented to the focus group participants is provided in Figure 5 on the next page. An example of the report format very similar in organization and structure (but not necessarily in content objectives) to the figure that was used in the focus group is shown in Figure 6 on page 38. The actual figure used in the focus group was based on a secure test form and remains confidential.

1. GRAPHICAL MAPPING OF ITEMS ON THE PACT TESTS (Item Map Graph)

- This approach uses a rectangular graphical display.
- The graph shows the relative difficulty of each item and the ability of students at the various cut scores.
  - The full range of scale scores, not just the cut scores, can be shown.
- The vertical axis represents the measurement scale reported in terms of scale scores.
- Both the difficulty of the items and the ability/achievement levels of the students are shown on the same scale using an IRT approach.
- The achievement level at each cut score is shown as a horizontal line across the page.
  - The achievement performance levels divide the scale into four groups, namely Below Basic, Basic, Proficient, and Advanced.
- The horizontal axis is used to locate the strands as nominal categorical variables.
- At each strand location, the items that measure content for that strand are shown by locating the items' positions on the vertical scale score dimension.
- A brief (one or two word) description of each item is provided on the map.
  - A more detailed content description can be provided on a separate page.

Figure 5. Description of the item content objective mapping strategy.

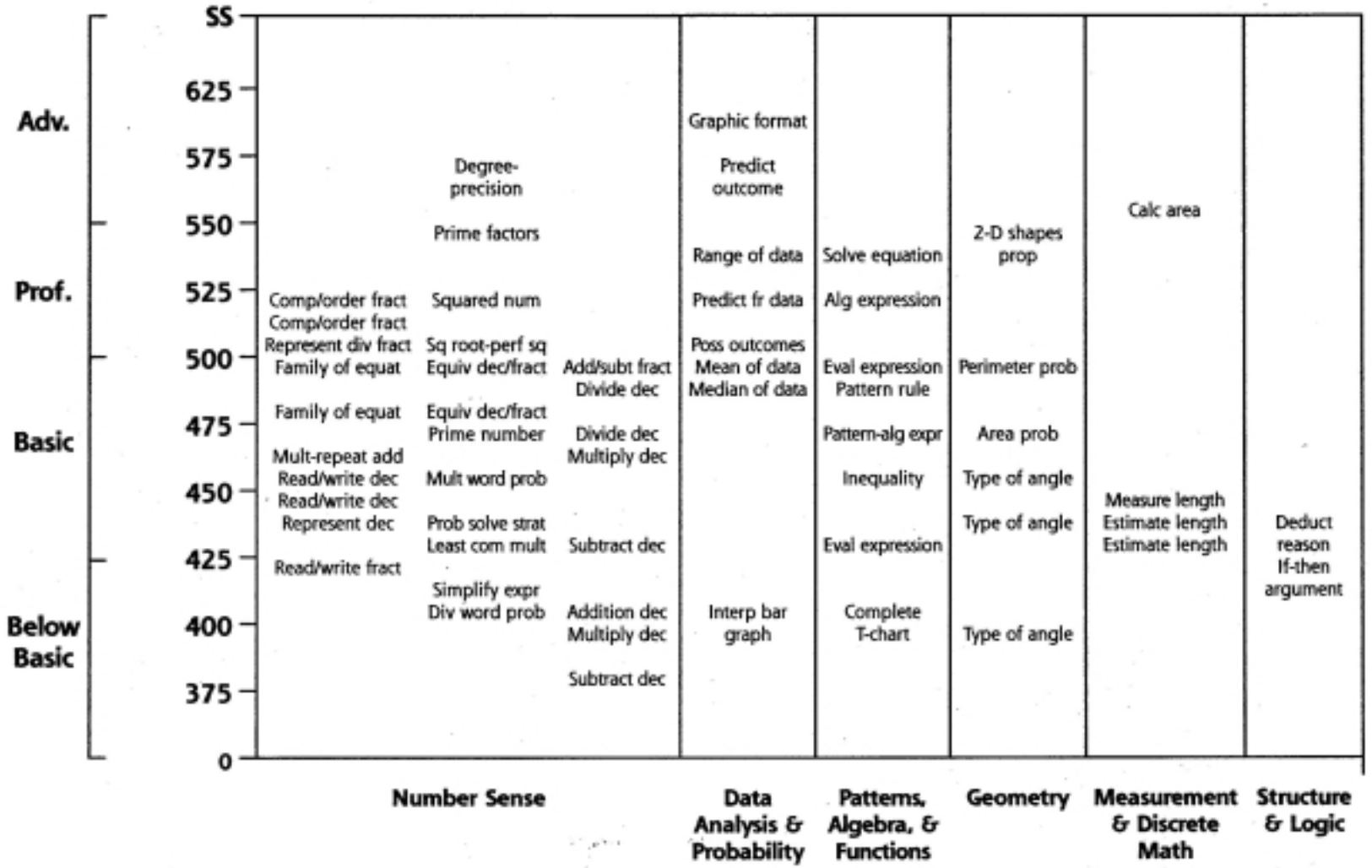

Figure 6. Example of an item content objective map for Grade 6, Mathematics

Focus Group Results: Item Content Objective Mapping

The comments of the focus group about this approach were quite mixed, with some participants offering strong endorsement, others indicating serious concerns about the approach, and still others offering some specific suggestions. The participants provided comments about pros, cons, and ways to improve the item mapping approach.

Strengths

The positive comments recorded by the participants on their worksheets and subgroup summary forms in response to the Item Content Objective Mapping strategy include the following.

1. *This is great!*
2. *I love the graph idea (as opposed to narrative).*
3. *Provides a ranking of scores from easy to difficult with specific reference to the cut points.*
4. *Could be valuable for determinations of staff development priorities*
5. *I like this because it tells you a lot about how the test works. Even if the questions change every year it gives you something to go on. It would help teachers predict how their students might do in certain areas.*
6. *Provides a content-based context for assessment interpretation.*
7. *Like [verbal] descriptions instead of curriculum codes.*
8. *Like the concept of associating scale score performance with strands/content.*
9. *Good at teacher level but need more descriptive information such as the narrative.*

These comments indicate that some of the participants believe that the item map or curriculum map approach is useful and can communicate assessment results effectively. There is a different opinion on this as will be seen shortly. The major substantive value seems to be that the assessment results are connected or embedded into the curriculum as represented by the strands and the item-level codes and descriptors. The connection between the cuts scores, scale scores, and content is also seen as a useful aspect of this score reporting approach.

Weaknesses

The areas of weakness recorded by the participants on their worksheets and subgroup summary forms in response to the Item Content Objective Mapping strategy include the following.

1. *Graphics difficult.*
2. *It is going to be hard to explain it. There will have to be an in-service for teachers.*

3. *Difficult to read, not for parents.*
4. *Requires loads of staff review – teachers and principals.*
5. *Good at teacher level but need more descriptive information such as the narrative.*
6. *Also, I would be concerned about explaining the 50-50 item chance to teachers and <u>especially</u> to parents.*
7. *Needs lots of description of what axes and indications mean (.50 response probability).*
8. *The major limitation of this form is that it does not encompass ALL the standards.*
9. *It may send the wrong message to teachers that only a part of the standards is important.*
10. *Big caveat – that this [map] changes from year to year.*

The issues expressed by the group address several topics. Participants were concerned with the use of a graph and the ability of teachers, parents, and perhaps others to correctly interpret graphically presented information. The group discussed (actually debated) the strategy of plotting items on the graph such that students had a .5 probability of answering an item correctly if their scale scores were on the same level as the item on the graph, e.g., if the students' ability and the item's difficulty were equal. A number of participants familiar with the NAEP achievement levels approach felt that a higher response probability would be more appropriate. Finally, the item map used in the focus group was based on a single field test form. Not all content standards are represented on each form and the group was concerned that the omission of some standards on a particular map might lead educators to misunderstand the need to direct instructional attention to all the eligible content standards.

Suggestions

The focus group participants had a number of suggestions about how the item map approach could be improved to make it more effective. These comments were not necessarily an endorsement of the item map approach if the suggestions were followed but were offered in the spirit that changes could be explored to refine the approach for future study.

1. *Add a narrative description of the levels and strands.*
2. *Provide a way to have students' names on the graph at the level that corresponds to their scale score.*
3. *Prepare appropriate professional development interpretive materials and training to accompany the use item maps. [This was implied in the explication of weaknesses.]*
4. *Add a cut line between Below Basic Level 1 and Below Basic Level 2.*
5. *If the map represents a single test form, list the standards that were not tested on that form.*
6. *List testable standards that aren't included in current year.*
7. *Consider 90% probability [as the response probability for graphing the items].*

8. *High-tech version [computerized report with "pop-up" item descriptor info] has appeal.*

The suggestions associated with changes in the item map and its uses are responsive to the identified weaknesses. The use of narrative descriptors and professional development materials and training speaks to the difficulties some felt teachers and parents would have in interpreting the item maps. The call for an additional cut line responds to the practice in South Carolina of subdividing the Below Basic category. The suggestion for a listing of standards that are not included in a given item map of a specific test form addresses the concern that educators might focus instruction too narrowly only on the content standards shown on a given map. The response probability issue is raised, with the suggestion of a map that shows "mastery" in the use of a response probability of .90. Finally, the reference to a high-tech version of the item map refers to a discussion of having item map score reports available electronically. Such maps would have an interface with several databases that could be accessed by clicking on the items on the map. The interface could provide such information as elaborated definitions of the content standard with examples, sample items that could be used to assess the standard, curricular and instructional resource materials, and connections to other useful web-based materials.

Discussion and Conclusions

The response of the focus group participants to the use of an item content objective mapping approach for reporting assessment results can be characterized as "shows promise, has some potential, but needs work." The value of such an approach seems to lie primarily in the way it integrates assessment results with content standards. Students' scores are shown on the same scale as the content standards and so the score has meaning by being embedded in the context of the content. The representation of the content objectives within achievement levels also reflects an alignment and connection between curricular standards and assessment performance.

The shortcomings of the graphical mapping approach seem to address three issues. First, what aggregation of items should be mapped? The map reviewed by the focus group represented a single test form, but several test forms or an entire bank of items could be plotted on the map. Care should be taken when using an item map that does not include items from all content standards that are eligible for assessment. In such cases, the map should have information that clearly indicates that other test forms would contain items measuring standards not represented on the current test and that all content specified in the state standards is eligible for assessment.

Second, what response probability should be used? A response probability of .50 was used in this study but in other applications, often a response probability of .67 or even higher is used. The choice of response probability is the subject of some discussion, as reflected in Section 2 of this report. Central to the discussion should be the questions of what information is being conveyed and how and by whom it is to be used.

- The response probability of .50 seems appropriate for information provided to teachers for the purpose of informing instructional decisions. In this context, the students' scores are aligned (literally on the map) with the content standards on which they demonstrate some knowledge but which they have not yet mastered.

- A response probability of .67 or higher would align students' scores with content objectives for which additional instruction might not be necessary since the students had already shown a level of attainment (.67 or higher) that suggests content mastery. In contrast, a response probability of .67 or higher seems appropriate for information provided to educators, parents and other stakeholders for the purpose of accountability and program auditing. If the purpose of the report is to classify students into various achievement levels and to track changes in the percentage of students at various levels, then the certainty with which students are classified is critical and a response probability of .67 or higher seems warranted.

A third issue raised in the focus group discussion was the need to provide adequate interpretive materials and professional development opportunities to help users of the reports in the assessment system. However useful the item content objective mapping approach may or may not be, materials and training support for any new reporting system are essential.

Finally, the development of the item content objective maps for this study yielded additional applications that might be useful to state departments of education and testing contractors. During the preparation of the item maps, a variety of options were explored including mapping one test, more than one test, or an entire item bank. These activities lead to the observations that the various maps provided a simple visual representation of the density of the items on a test, test forms, or item bank, relative to the content strands and achievement levels.

- An item map can be used during the development of an operational form to examine how many items in each strand are located within each achievement level. Items might then be "switched" in and out to construct a test form with items from all strands represented at each achievement level.

- The comparison of an item map for a test form compared to the item map for the test bank provides a useful visual display of how well the test form reflects the bank.

- When additional item development is scheduled, an item map of the entire item bank can be used to identify content strands and achievement levels for which item development might be targeted. The procedure described here for monitoring tests and item banks is currently done analytically using IRT logit values and the use of the item maps is suggested as a complementary strategy.

In conclusion, the item content objective mapping approach explored in this study seems to have promise as a reporting format but additional development work is necessary before such an approach can be considered for implementation.

B2. Achievement Performance Level Narrative

Introduction

The achievement performance-level narrative approach involves placing each item from a test into one of the four achievement-level categories based on the item's difficulty and then developing a narrative that describes the content and content demands at each achievement level. As mentioned in the literature review, there is considerable discussion in the psychometric community about exactly which items should be considered exemplars and included as part of the narrative for the respective achievement levels. The concern is that students in the lower end of an achievement level might have a relative low probability of correctly answering items from the upper end of the achievement level (p<.50).

In this part of the study, the items were reviewed from all test forms that had been given in the current program in Grade 3, Mathematics and Grade 8, English/Language Arts. To develop the performance-level descriptions in the narrative, items were selected if the probability of students at the cut score correctly answering the item was approximately .67, generally ranging from .60 to .75. This response probability range was used because fewer exemplars would have been identified if a more restricted response probability were employed. In addition, an item was not used as an exemplar unless at least one other item measuring the same content objective was also eligible as an exemplar. This requirement was added so that no part of the achievement-level description would describe content objectives represented by a single item. Such an approach would too strongly resemble describing the test items as opposed to general characteristics of the content objectives.

A content panel looked at items with lower and higher response probabilities for the strands and objectives in order to get a better sense of the substantive features involved. The staff in the Office of Assessment prepared the item data, and state department content experts prepared the achievement-level descriptions.

Description and Explanation of the Reporting Strategy and Format

The information on the achievement-level performance narrative strategy that was presented to the focus group participants is provided in Figure 7 on the next page. The achievement-level descriptions generated by this process and used in the focus group are shown in Figures 8 and 9.

2. NARRATIVE DESCRIPTION OF ACHIEVEMENT LEVEL PERFORMANCE

- Each item that had been used on a PACT test was placed into one of the four achievement-level categories based on the item's difficulty.
- The expected proportion of students at the cut score for each achievement level answering each question correctly was calculated.
- For mathematics, items were sorted by strand and were then ordered from easy to hard within each strand.
  - A panel of mathematics content experts studied the items in each achievement level within each strand and developed a narrative describing what knowledge and skills students answering these items would be demonstrating.
  - The panel examined the text of each item as well as the proportion of students at each cut score expected to answer the item correctly.
  - Content elements represented by only one item were not described because a "generalized" description could not be constructed from a single example.
- For English Language Arts (excluding writing), items were arranged from easy to hard within the set of items related to each passage.
  - A panel of ELA content experts studied the items in each achievement level within each strand and developed a narrative describing what knowledge and skills students answering these items would be demonstrating.
  - The panel examined the text of each item as well as the proportion of students at each cut score expected to answer the item correctly.
  - Content elements represented by only one item were not described because a "generalized" description could not be constructed from a single example.

Figure 7. Description of the achievement performance-levels narrative strategy.

Third Grade Mathematics Achievement Level Descriptions

Below Basic

Third-grade students scoring at the "Below Basic" level are able to estimate and perform basic operations with whole numbers. They can identify simple number sentences and expressions, simple patterns, and common two-and three-dimensional geometric figures and geometric properties. The student at the "Below Basic" level can read tables and answer questions based on data contained in the tables as long as the questions require no more than simple computations.

Students scoring at the "Below Basic" level tend to be unable to solve multi-step problems and problems involving division. They tend to be weak in measurement. They also tend to have difficulty reading and interpreting scales and working with pictorial representations.

Basic

Third-grade students scoring at the "Basic" level are able to answer problems requiring more than one-step or operation, alternate between two different types of patterns, and apply straightforward concepts of probability. Their performance differs from the performance of students scoring at the "Below Basic" levels in the amount of data that can be handled, the number of steps required by the problem, the nature of the mathematics vocabulary, and the degree of reasoning required.

Student scoring at the "Basic" level do not appear adept with measurement concepts such as reading and interpreting scales. They also have difficulty working with pictorial representations, fractions, and division.

Proficient

Third-grade students scoring at the "Proficient" level are able to interpret and translate pictorial representations. They exhibit an understanding of the concepts of fractions and division. They can apply straightforward measurement concepts. When units of a scale are marked, they are able to read and interpret scales. Students scoring at the "Proficient" level are able to translate language into numerical concepts.

Students scoring at the "Proficient" level tend to have difficulty problem solving when required to use spatial sense.

Advanced

Third-grade students scoring at the "Advanced" level make connections among mathematic ideas and communicate their mathematical thinking and reasoning coherently and clearly. They have stronger spatial sense than other students. They are more tenacious than students at other levels in approaching problems that appear longer and/or more complex. They are able to tackle problems requiring approaches that are not commonly used.

Figure 8. Example of an achievement performance-levels narrative for Grade 3, Mathematics.

Eighth Grade English/Language Arts Achievement Level Descriptions

Below Basic

Eighth-grade students scoring at the "Below Basic" level are able to skim and locate obvious details using key words or phrases in passages that are of high interest to them. When the passage provides a stated main idea, the student at the "Below Basic" level is able to identify that main idea, and he or she is able to draw simple conclusions about the passage when the text provides obvious support for those conclusions.

Eighth-grade students scoring at the "Below Basic" level tend to be unable to locate details in longer, denser passages. They tend to be unable to handle poetry, and they are unable to combine reading strategies in order to draw higher-level conclusions about the text they read.

Basic

Eighth-grade students scoring at the "Basic" level are able to locate details in longer passages and make simple inferences from informational and literary text that is of high interest. They are able to paraphrase the main idea, and they are able to provide literal interpretations in reading informational and literary text. Students scoring at the "Basic" level are able to combine strategies (e.g. locate details to make an inference) while reading, and they are able to recognize the literary elements (e.g., simile and point of view) first introduced during elementary school.

Eighth-grade students scoring at the "Basic" level tend to have difficulty providing literal interpretations for poetry. They tend to have difficulty analyzing literary elements and figurative language introduced in middle school. They also tend to have difficulty in going beyond the text to answer constructed response questions or supporting their response with details.

Proficient

Eighth-grade students scoring at the "Proficient" level are able to make distinctions among and analyze details to make more complex inferences regarding the longer, denser informational, literary, and poetic text that they read. Eighth grade students scoring at the "Proficient" level are able to understand and analyze both literal and figurative language, and they are adept at interpreting and drawing conclusions in poetry. They are able to go beyond the text to answer constructed response to questions and tend to support their responses with details.

Eighth-grade students scoring at the "Proficient" level tend to have trouble evaluating reading material, and their written responses, while accurate, tend not to be insightful and creative.

Advanced

Eighth-grade students scoring at the "Advanced" level are able to make fine distinctions among many details to make more complex inferences regarding the longer, denser informationally, literary, and poetic text that they read. They are able to understand, analyze, and evaluate both literal and figurative language, and they are adept at interpreting and drawing conclusions in poetry. In addition, advanced students are able to provide detailed, complete, insightful, and creative answers to constructed questions relating to written text.

Figure 9. Example of an achievement performance-levels narrative for Grade 8, English/ Language Arts.

Focus Group Results: Achievement Performance Level Narrative

Strengths

The positive comments recorded by the participants on their worksheets and subgroup summary forms in response to the narrative descriptions of achievement levels include the following.

1. *Good. Need for all areas ASAP.*
2. *Parents would love it.*
3. *Good for parents.*
4. *Good information for parents and communities.*
5. *Good communication tool for parent conferences.*
6. *Yes, teachers will find useful. Teachers could use these narratives in talking with parents. Principals could use this in conjunction with test scores to focus on possible curriculum alignment and implementation in collaborative planning.*
7. *This information is quite useful for parents or general public. Based on PACT items throughout these years, content experts come up [with] summary like this and I think it's great!*
8. *Good document for principals and teachers.*
9. *In general, this is the most useful document for me as a principal.*
10. *Like labels instead of codes.*

The written comments in response to the Achievement Level Descriptions were clearly quite positive, as were the oral comments during the discussion. The participants felt very strongly that this approach would be useful, informative, and helpful for teachers and principals and especially helpful in communicating with parents and other members of the community.

Weaknesses

The areas of weakness recorded by the participants on their worksheets and subgroup summary forms in response to the achievement-level performance descriptions approach included the following.

1. *Not specific enough – need to pull levels apart – which things can my kids not do?*
2. *Doesn't pass "so what" test.*
3. *Limited to areas where pattern can be identified.*

These are relatively minor concerns in that more specificity is generally desired. The lack of patterns refers to the decision not to include an item if only one item from an objective was eligible.

Suggestions

The focus group participants had a number of suggestions about how the Achievement Performance Level Narrative strategy could be improved to make it more effective.

1. *Bulleted format; Bullets instead of narrative; Make these in bullet form; Put in bullet form; Reformat bulleted list, matrix format would be better; Better in a bullet form. Make clearer in format - bullets, etc.; Bullets of matrix would be easier to read.* [This is a list from all participants. There were eight suggestions to reformat the narrative into bullet form].
2. *Would be nice to have supporting instructional documents.*
3. *Need supporting documents/ best practices.*
4. *Supportive documents (best practices).*
5. *Indicate score range for levels.*
6. *Put score range in each category.*
7. *Scale score ranges.*
8. *BB1 and BB2.*
9. *Add BB1.*
10. *BB1 and BB2 and descriptions.*
11. *List areas for which there was no pattern.*
12. *What about the areas where no patterns were established? I would like to see another listing of areas where there were no patterns.*
13. *For all the suggested formats, I would strongly suggest making it electronic.*
14. *Electronic version on web site.*
15. *Clarify in writing what cut point definitions are – just barely above cut point.*
16. *Would like to combine concept with information for #6 [Item-analysis information].*
17. *Teachers would need to understand that any year's test would not be a perfect match.*
18. *Could this type/format of information be available in addition to the graphical presentation?*
19. *Need to pinpoint by strands to be helpful to educators.*
20. *Combine with graphical map.*
21. *Add disclaimers – any single year's test may not be perfect match.*

The focus group participants were quite positive about the narrative description of achievement performance but, nevertheless, had a number of suggestions for improving this approach. The most frequent suggestion was to take the narrative and present it in bullet form.

The focus group members discussed the value of connecting instructionally related materials to each achievement level so that teachers and others could use the information in a diagnostic-prescriptive fashion. The groups suggested that separate descriptions of the two levels of Below Basic be provided and that the score ranges be added to the narrative descriptions.

The group recognized that not all objectives would be represented in the description and referred to such objectives as showing "no pattern." They suggested providing a list of objectives that were tested but not represented in the narrative. In addition, there was a discussion about the value of having these descriptions available electronically.

Discussion and Conclusions

Focus-group members strongly endorsed the use of the narrative descriptions of achievement-level performance as a useful and informative approach to reporting student test results. Despite support for a narrative approach, the group suggested that the narrative be deconstructed into bulleted points. As an example of this, the description of Proficient Level for third grade mathematics in bullet form is shown below.

Third grade students scoring in the Proficient level:

- *Are able to interpret and translate pictorial representations.*
- *Exhibit an understanding of the concepts of fractions and division.*
- *Can apply straightforward measurement concepts.*
- *Are able to read and interpret scales when units of a scale are marked.*
- *Are able to translate language into numerical concepts.*
- *Tend to have difficulty problem solving when required to use spatial sense.*

A comparison of the bulleted format and the narrative format supports the focus-group members' comments that the bulleted format seems to make the information clearer and easier to understand.

In conclusion, the achievement-level performance description approach seems to be effective and was strongly recommended for broad implementation as soon as feasible. The value of using a bulleted format needs to be considered.

B3. Strand Achievement Levels for Individual Students

Introduction

The first focus group suggested that reporting students' achievement at the strand level in terms of the achievement-level classifications of Below Basic (BB), Basic (B), Proficient (P) and Advanced (A) would be useful for educators. Such an approach would provide a finer level of detail that is often thought to have more diagnostic value than classifying students on the basis of the total test performance. All standardized norm-referenced tests results are reported at a level of detail finer than the total test score and educators seem accustomed to having such information.

Developing strand-level achievement reports involves interpolating cut scores from the total test level to each strand. In this study, the procedure used the IRT *theta* at each cut score and the item difficulties of the items on each strand for an operational test form. The probability of a correct response to each item for students at the cut score (*theta*) was calculated using the one-parameter IRT model. With the one-parameter IRT model, the sum of these probabilities is the expected raw score on the strand for students at the cut score being evaluated.

Students who take this test have an observed number of items answered correctly on each strand and this observed strand-level score can be compared to the expected strand-level cut score to classify the student. Generally, the expected strand cut scores are not integers so the interpolated values are rounded to the nearest integer value.

Description and Explanation of the Reporting Strategy and Format

The information on the strand achievement-level strategy that was presented to the focus group participants is shown in Figure 10. The example of a report based on this procedure that was used in the focus group involved a small sample of data with student names removed. The report is shown as Figure 11.

3. STRAND LEVEL INTERPRETATION OF ACHIEVEMENT LEVELS FOR INDIVIDUAL STUDENTS

- Cut scores for the achievement levels on the total test are used to estimate performance levels on each strand so that each strand may have Below Basic, Basic, Proficient, and Advanced levels.
  - Estimated strand-level cut scores are not integer values so they are rounded to the nearest integer.
  - The rounding of the strand-level cut scores introduces some error.
- Students' scores on each strand are calculated.
- Students' observed strand-level scores are compared to the strand-level cut scores and students are classified into Below Basic, Basic, Proficient, and Advanced on each strand.
- For some strands it is not possible to attain a level of "Advanced" because there were not a sufficient number of items at the Advanced level on that strand.
- Great care must be used when interpreting students' achievement levels on the strands.
  - The standard error of measurement is large for the strands because of the relatively small number of items used on each strand.
    - (Most differences between strand-level performances for individual students are not statistically significant.)
  - Students' achievement-level classification on the total test is NOT a simple sum or average of students' achievement levels classifications on the strands.
  - Students with the same achievement level based on the total test could have different profiles of achievement-level classifications on the strands.

Other sources of information about individual student's achievement, such as classroom assessments, samples of students work, and cumulative records, etc., should be used along with strand-level achievement levels when evaluating a student.

Figure 10. Description of the strand achievement levels for individual students' strategy.

Student ID	Total Test	Strand 1	Strand 2	Strand 3	Strand 4	Strand 6
		Number & Operations	Algebra	Geometry	Measure-ment	Data Analysis & Probability
	Level	Level	Level	Level	Level	Level
XXXXXX	BB	B	B	BB	BB	BB
XXXXXX	BB	BB	BB	BB	BB	BB
XXXXXX	B	B	B	B	BB	BB
XXXXXX	BB	BB	BB	BB	BB	BB
XXXXXX	BB	BB	BB	B	BB	BB
XXXXXX	B	B	B	P	BB	BB
XXXXXX	BB	BB	BB	BB	BB	BB
XXXXXX	B	B	P	B	A	B
XXXXXX	BB	BB	BB	B	BB	BB
XXXXXX	B	B	B	BB	BB	B
XXXXXX	P	B	P	P	A	B
XXXXXX	P	P	P	P	A	B
XXXXXX	BB	BB	BB	B	BB	BB
XXXXXX	B	B	P	B	BB	BB
XXXXXX	P	B	B	A	A	P
XXXXXX	B	B	P	P	BB	B
XXXXXX	P	A	B	B	A	A
XXXXXX	B	B	P	B	BB	B
XXXXXX	B	B	BB	B	BB	BB
XXXXXX	BB	BB	P	BB	BB	BB
XXXXXX	B	B	B	B	BB	P
XXXXXX	A	A	P	A	P	P
XXXXXX	P	P	P	P	P	A
XXXXXX	BB	BB	BB	BB	BB	BB
XXXXXX	B	B	P	B	B	BB
XXXXXX	BB	BB	BB	BB	BB	BB

Figure 11. Example of a strand achievement-level report for individual students for Grade 3, Mathematics.

Focus Group Results: Strand Achievement Levels for Individuals Students

Strengths

The positive comments recorded by the participants on their worksheets and subgroup summary forms in response to the strand achievement-level report for individual students include the following.

- *Despite the cautions this would be useful information to have school and district wide.*
- *Parents/teachers would like this.*
- *Parents and teachers would think it easy to read.*
- *Good for student-level data.*
- *Quick and dirty.*
- *It would be great if information were more reliable.*
- *Might have some value at the district level where lots of data are available.*
- *Good information butt not enough for true instructional decisions.*
- *Like seeing pattern of individual strengths and weaknesses.*

There was clearly some positive response to reporting students' achievement-level classification at the strand level. Several focus group participants felt that this kind of information might be appropriate for teachers and parents. Some positive comments, however, were qualified.

Weaknesses

The areas of weakness recorded by the participants on their individual worksheets and subgroup summary forms in response to the strand achievement-level interpretation for individual students include the following.

- *I'm not sure how useful it would be to an individual teacher.*
- *I would not be comfortable sharing this with teachers.*
- *May be confusing to teacher.*
- *Teachers might make false instructional or placement decisions.*
- *Doesn't impact instruction.*
- *Good information but not enough for true instructional decisions.*
- *Poor conclusions.*
- *Too much room for bad decisions on this one.*
- *Too much room for misinterpretation.*
- *Misleading information.*
- *Not enough information.*

- *Information too broad.*
- *Too general.*
- *Confusing*
- *Confusing (averaging B's, P's, A's [Basic, Proficient, Advanced])*
- *As is, this is not helpful if the reliability is questioned.*
- *It would be great if information were more reliable.*
- *Unreliable*
- *Not good*
- *Don't use this*

The focus-group participants clearly expressed a number of concerns about reporting individual student's achievement levels for each strand. The concerns seem to focus on 1) its value (or lack of) for teachers making instructional decisions; 2) the general nature of the information as being broad, misleading, and perhaps confusing; and 3) the lack of reliability in the classifications at the strand level. The concern expressed about the reliability may reflect cautionary comments made in the presentation to the focus group. These comments about reliability may have influenced the group's view of the usefulness of this information to teachers.

Suggestions

The focus-group participants had a number of suggestions about how the strand-level reporting of the achievement levels could be improved to make it more effective.

- *One would have to look at it over time.*
- *Could you provide the information from Format 2 [narrative] by strand? For all years of PACT?*
- *Narrative gives information and would also give a level.*
- *If we ask for number of items in each strand being printed on the chart, it might help the reader interpret the scores a little better if this type of summary is to be provided.*
- *Use student names not ID.*
- *Perhaps use this for classrooms, but not students (too precarious statistically).*
- *A place for individual students' <u>names</u> (not just ID#).*

The suggestions of the participants about this reporting approach were not particularly substantive but seemed to address some general features of the approach.

Discussion and Conclusions

Focus-group participants had some mixed thoughts about the use of strand-level classifications but generally seemed to express reservation about this approach. The concerns expressed by the focus group about the reliability of the strand-level reports may reflect the suggestions of the facilitator during the exposition to the group on this method.

The concern about this approach generally relates to the lack of reliability at the strand level. The strands have a small number of items and therefore likely to lack the reliability recommended for making inferences about students. The reliability of the strands was explored in detail and is presented below.

Investigation of the Reliability of Strand Level Subtests

Measurement theory and practice show that the reliability of tests with small numbers of items can be low and questionable. The facilitator conveyed this concern when presenting the strand achievement levels strategy for individual students to the focus group participants. Their response to this approach also echoed this concern. To determine the extent of the reliability of the strands, an empirical investigation was conducted and the results of this investigation are presented for Grade 3, Mathematics strands and Grade 8, English/ Language Arts areas in the remainder of this section.

Mathematics Strands

The comparison of strand-level performance requires that strand achievement be on the same scale for all strands. This was accomplished by using the test-level IRT difficulty values for the items of each strand to estimate a strand-referenced IRT ability. Table 5 shows the number of points on each strand and the mean and standard deviation in the common IRT logit scale. The reliability of each strand is shown, and ranges from .44 for Data Analysis & Probability; approximately .60 for Algebra, Geometry, and Measurement; and .83 for Number & Operations. Reliabilities in the .40s are certainly lower than required for making important decisions about students and reliabilities in the .60s are also questionable.

Table 5

Descriptive Statistics and Reliabilities for Grade 3, Mathematics Strands
(IRT Logit Scale)

	Subscale	Number of Items/Points	Mean	SD	Reliability
1.	Number & Operations	10	0.05	1.55	0.83
2.	Algebra	8	0.08	1.36	0.64
3.	Geometry	10	0.04	1.35	0.60
4.	Measurement	7	-0.09	1.59	0.58
5.	Data Analysis & Probability	5	-0.04	1.54	0.44

In addition to the reliabilities of the strands, it is important to know the correlations of students' performance among the strands. These correlations provide some information about the degree to which the information in the strands is unique or shared across the strands. The intercorrelations among the Grade 3, Mathematics strands are reported in Appendix B.

The major application of strand (or subscale) results is to make inferences about students' relative strengths and weaknesses. Educators and others want to use such information to describe student performance. For example, a student might be described as doing relatively better in Number & Operations, Measurement, and Geometry, than in Algebra and Data Analysis & Probability. Such comparisons are inferences about difference in students' performance on the various strands and the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, and National Council on Measurement in Education [NCME], 1999) require that the reliability of such an inference be reported. The reliabilities for the differences in performance on the strands are calculated from the strand reliabilities (see Table 5) and the intercorrelations of the strands (see Appendix B). The reliabilities of the differences between strands are presented in Table 6.

Table 6

Reliabilities of Strand Differences for Grade 3, Mathematics (IRT Logit Scale)

Subscale	2. Algebra	3. Geometry	4. Measurement	5. Data Analysis & Probability
1. Number & Operations	0.35	0.34	0.24	0.24
2. Algebra		0.16	0.08	0.10
3. Geometry			0.04	0.03
4. Measurement				-0.06

The range of reliabilities of the differences in students' performance on the strands is from -.06 to .35. These reliabilities are well below the level of what is generally acceptable for making inferences about students and suggest that claims about students being stronger or weaker in various strands are based on differences that are generally not much greater than random variation.

To determine the standard error (SE) of the differences, the mean and standard deviation for the differences for all pairs of strands were calculated. The standard deviations of the differences and the reliabilities of the differences were then used to calculate the standard errors of the differences between strands. These standard errors are reported in Table 7.

Table 7

Standard Errors of the Strand Differences for Grade 3, Mathematics (IRT Logit Scale)

Subscale	2. Algebra	3. Geometry	4. Measurement	5. Data Analysis & Probability
1. Number & Operations	1.04	1.06	1.21	1.32
2. Algebra		1.18	1.31	1.41
3. Geometry			1.33	1.44
4. Measurement				1.59

A 95% confidence interval (CI) around each difference can be constructed (1.96 x the SE) to identify the range of scores for which the differences do not exceed chance level. For example, the comparison of students' performance on Number & Operations to Algebra has a 95% confidence interval of $\pm$ 2.04. This confidence interval is used to evaluate the statistical significance of the differences between students' performance on the two strands. Each student in the state database for 2002 (N> 48,000) has ability estimates based on raw scores from the Number & Operations and Algebra strands and thus, for each student, the difference in their performance on the two strands can be calculated. The distribution of the differences for the population of test takers was compared to the 95% confidence interval for Number & Operations compared to Algebra and only 12% of the students had differences that fell beyond the 95% confidence interval. The results for all pair-wise comparisons across the strands are shown in Table 8.

Table 8

Percent of Students with Statistically Significant Differences in Their Strand Level Performances

Strands Compared	SE	95% CI($\pm$)	% Ss Beyond the 95% CI
1-2	1.04	2.04	12.00
1-3	1.06	2.08	10.39
1-4	1.21	2.37	7.87
1-5	1.32	2.59	8.80
2-3	1.18	2.31	7.34
2-4	1.31	2.57	4.19
2-5	1.41	2.76	4.30
3-4	1.33	2.61	4.08

3-5	1.44	2.82	4.56
4-5	1.59	3.12	3.89

The use of the 95% confidence interval for interpreting these comparisons treats each pair of differences as an independent comparison. These 10 pair-wise comparisons among differences of correlated strands are not independent, however. A more conservative approach to control for Type I error would be to distribute the .05 across the ten comparisons by using .05/10 or .005 as the significance level for each comparison. The overall error rate for the 10 comparisons would be .05 in such case. This amounts to using a 99.5% confidence interval to interpret the differences, and at this confidence interval level fewer than 2% of the differences between students' performance on pairs of strands are significantly different.

A final analysis conducted for Grade 3, Mathematics test data examined the consistency in the classification of students at the strand level. The *kappa* indices of agreement in classification based on a single test administration (Huynh, 1976, 1978) were determined, and the data from this investigation are reported in Appendix C. The results show a very low level of agreement and consistency in the classification of students into achievement levels across the strands.

Grade 8, English/Language Arts Areas

Similar analyses with Grade 8, English/Language Arts areas were conducted and the results lead to the same conclusions. Table 9 shows the number of points for each area and the mean and standard deviation in the IRT logit scale.

Table 9

Descriptive Statistics and Reliabilities for Grade 8, English/Language Arts
(IRT Logit Scale)

ELA Areas	Number of Items/Points	Mean	SD	Reliability
1. Communication	6	0.10	1.48	0.47
2. Research	6	0.11	1.36	0.30
3. Reading	60	0.02	1.00	0.88
4. Writing	27	0.09	1.35	0.80

The reliabilities of the difference scores for Grade 8, English/Language Arts areas are shown in Table 10.

Table 10

Reliabilities of the Area Differences for Grade 8, English/Language Arts (IRT Logit Scale)

ELA Areas	2. Research	3. Reading	4. Writing
1. Communication	0.05	0.13	0.29
2. Research		0.08	0.22
3. Reading			0.43

The standard errors for the area differences are shown in Table 11.

Table 11

Standard Errors of the Area Differences for Grade 8, English/Language Arts (IRT Logit Scale)

ELA Areas	2. Research	3. Reading	4. Writing
I. Communication	1.57	1.13	1.23
2. Research		1.19	1.29
3. Reading			0.70

And finally, Table 12 presents the percentage of students beyond the 95% confidence interval that would indicate a statistically significant difference in the performance on the four English Language Arts areas. As can be seen in this table, only 5% to 12% of the students have statistically significant differences in their strand-level performance. In all other cases, differences that might be reported as indicating relative “strengths” and “weaknesses” across the strands are no greater than random variation.

Table 12

Percent of Students with Statistically Significant Differences in Their Area Level Performance

Areas Compared	SE	95% CI(±)	% Ss Beyond the 95% CI
1-2	1.57	3.08	5.25
1-3	1.13	2.21	6.44
1-4	1.23	2.41	9.96
2-3	1.19	2.33	6.42
2-4	1.29	2.53	7.27
3-4	0.70	1.37	12.08

These six pair-wise comparisons among differences of correlated strands are not independent, however, and a more conservative approach to control for Type I error would be to distribute the .05 across the six comparisons with the result that virtually all the differences between students' performance on pairs of strands would be no greater than random variation.

Conclusions

In conclusion, the caution about the reliability of the strands and comparisons between strands expressed in the presentation to the focus-group participants and the views of the participants during the focus-group discussions are supported by the subsequent data analyses. The differences in students' performance across the strands are not reliable enough to be used to make inferences about differences in students' relative abilities and achievement across strands and should not be reported at the level of individual students.

B4. Strand Achievement Levels for Groups

Introduction

This approach to reporting PACT scores provides information about students' performance on each strand based on grouped data and, therefore, results are reported in percentages. This is similar to the previously described approach but aggregates the data by groups such as schools, districts, and state.

Some of the same cautions as described for the previous approach apply to the group-based approach, but the group statistics are more stable since they reflect averages and the errors of measurement are likely random and would cancel out.

Description and Explanation of the Reporting Strategy and Format

The information about the strand achievement levels strategy for groups of students that was presented to the focus-group participants is shown in Figure 12 below. An example of the report format based on this approach is provided as Figure 13.

4. STRAND LEVEL INTERPRETATION OF ACHIEVEMENT LEVELS FOR GROUPS, SUCH AS SCHOOLS, SCHOOL DISTRICTS, AND THE STATE

- Strand-level achievement classifications for students are determined as described in the preceding procedure (#3).
- A summary table is constructed showing the percentage of students in the Below Basic, Basic, Proficient, and Advanced levels based on the total test and for each strand.
- For some strands it is not possible to attain a level of "Advanced" because there were not a sufficient number of items at the Advanced level for that strand.
- Care must be used when interpreting the percentage of students in each achievement level for each strand because of the small number of items used to assess each strand.

Figure 12. Description of the strand achievement levels strategy for groups.

STRAND LEVEL INTERPRETATION OF ACHIEVEMENT LEVELS FOR GROUPS, SUCH AS SCHOOLS, SCHOOL DISTRICTS, AND THE STATE

Achievement Level	Total Test	1 Number & Operations	2 Algebra	3 Geometry	4 Measure-ment	5 Data Analysis & Probability
	%	%	%	%	%	%
Advanced	5.90	8.12	*	8.89	15.93	7.64
Proficient	13.44	11.43	27.13	13.89	13.92	18.77
Basic	34.12	41.02	23.90	32.93	16.48	26.10
Below Basic	46.55	39.43	48.97	44.29	53.68	47.49

**There were no Algebra items of sufficient difficulty to allow students to demonstrate advanced status on the strand.*

Figure 13. Example of a strand achievement-level report for groups

The sample chart used to explain this reporting approach to the participants reflects the fact that for some strands, e.g., Algebra, there may not be items that allow students to demonstrate the highest level of performance.

Focus Group Results: Strand Achievement Levels for Groups

Focus-group participants discussed the strengths and weaknesses of this approach and their comments written on the individual information sheets, the group summary charts, and oral comments revealed a number of questions. In addition, a number of written comments include both a positive observation and a statement about a weakness or limitation. Such comments are repeated under both the "Strength" and "Weakness" headings.

Questions about the Approach

The group raised the following questions about this approach.

- *What size group [is needed] for adequate confidence level?*
- *How big does the group have to be for the percentages in the different strands to be accurate enough to make curriculum decisions? Class? Grade level? School? District?*
- *Would this be valid for reporting classroom data? (Maybe have a confidence interval chart based on number of students along with it.)*
- *Is this reliable at the class level?*
- *For small school districts with only one school per grade span would this be good information?*

These questions all focus on the concern about how large a sample would be needed to make accurate inferences about students' performance on the different strands. This concern probably reflects the previous discussion about the unreliability of inferences about individual student's performance at the strand level.

Strengths

The positive comments recorded by the participants on their worksheets and subgroup summary forms in response to the strand achievement-level approach for a group of students include the following.

- *I think a school could really benefit from this information if it works for a grade level (300 students). If not classroom, do for grade level.*
- *OK for gross patterns but not helpful for flexible grouping within classes.*
- *OK for district. No instructional data except in broadest sense.*
- *Helpful for districts and large schools*
- *Useful only at the district level.*
- *Best for district use.*
- *Helps identify instructional weaknesses that are generalized across classrooms.*
- *I think we ought to recognize that PACT is really an accountability system; therefore, using PACT results to pinpoint/determine instructional strengths/weaknesses on an individual level is not a perfect solution. The Group Level summary seems to me the most appropriate. Use it with Blue Prints.*
- *Good for district level, but not specific enough for teachers to identify specific instructional purposes.*
- *Interesting to compare school-to-school, grade to grade. Not too reliable for class to class, however teachers and principals would love this!*

The feedback from the group about this approach suggests a certain value in monitoring students' strengths and weaknesses relative to the curriculum strands for relatively large aggregations of students, such as school districts, schools, and possibly grade levels in large schools. These comments also reflect a concern that the information would not be very useful with small groups of students.

Weaknesses

The areas of weakness recorded by the participants on their individual worksheets and subgroup summary forms in response to the strand-level interpretation for groups of students include the following.

- *OK for gross patterns but not helpful for flexible grouping within classes.*
- *OK for district. No instructional data except in broadest sense.*
- *Not too reliable for class to class …*
- *Not useful for small school district; not good at school level.*
- *Must have sufficient group size.*
- *Not useful for small district and not good at school level.*
- *Broad paintbrush*
- *Limited utility at classroom level. Won't help identify students for flexible grouping.*
- *Not reliable for principals at school level.*
- *Good for district level, but not specific enough for teachers to identify specific instructional purposes.*

The comments of the group indicate that the information at the strand level is still too broad to provide teachers and others with specific information about strengths and weakness even for a group. In addition, the comments suggest a belief that the approach is viable in terms of reliability for only relatively larger aggregations of students such as school districts, possibly schools, and grade levels in large schools.

Suggestions

The focus group participants had a number of suggestions about how the strand-level reporting for groups might be adapted.

- *If we use this design, we might want to consider adding difficult items in algebra.*
- *Interesting to compare school to school, grade to grade*
- *Need to do by sub-populations (true for all forms).*
- *MAP [a commercial testing program] testing does this and it helps focus instruction in schools.*

Discussion and Conclusions

The response of the focus group to reporting strand performance of groups of students for various achievement levels is mixed. Positive sentiments suggest a value in examining relative strengths and weaknesses in students' learning for the different strands. This was seen as useful and likely to be well received. However, there was concern that to be reliable the data would have to be based on relatively large groups of students and the larger the group, the less helpful the data might be. This reporting format was seen as most likely to be helpful at the district level and to be of less help at the classroom level and even grade level for small schools.

B5. Observed, Expected, and Differences in Strand and Item Performance for a Group

Introduction

The approach to score reporting examined here is based on the fact that each district and school has a mean scale score on each PACT test that reflects a mean Item Response Theory (IRT) ability. This mean IRT ability and the IRT item difficulties can be combined in the IRT model to estimate the probability that students with that given mean ability will correctly answer each item on the test. This probability can be viewed as a predicted proportion as compared to the observed proportion of students answering each item correctly. An example of the score report format for this strategy is provided in Figure 15.

This approach can be interpreted as reporting how well a district is doing on each item using the district's overall (mean) performance as a control. In some respects, this type of report would serve some of the same purposes as a traditional item-analysis report. In the South Carolina assessment program, an item-analysis report had been provided to schools previously but is not part of the current PACT reporting procedure. The use of an item-analysis report, or close approximation of one, would be a return to a score reporting procedure that some South Carolina educators found useful.

It is important to note that the items in the example shared with the focus group were not listed in their order on the test. Rather, the items were grouped by strand. In addition, strand-level means for each column are reported. The level of detail in the item descriptions is critical but was left blank in the example.

Providing "Interpretation Guidelines" for users of this type of report is important given the somewhat complex nature of the type and amount of information provided. Guidelines about "how big" a difference (percentage above or below the expectation) should be in order to be considered serious would need to be included to help educators interpret the information appropriately.

Description and Explanation of the Reporting Strategy and Format

Information about the observed and expected differences of strand/item performance strategy that was presented to the focus-group participants is provided in Figure 14.

5. GROUP OBSERVED PERFORMANCE COMPARED TO EXPECTED PERFORMANCE ON EACH ITEM

- Each district and school has a mean scale score, which reflects a mean ability or mean achievement measure on the underlying scale.
- The group's mean ability or achievement measure is used to predict what proportion of students in the group is expected to correctly answer each item.
- The observed proportion of students in a group correctly answering each item is calculated.
- The difference between the observed and expected proportion of students correctly answering each item is calculated and reported by strand.
- For the items on each strand, the means for the observed and expected proportion of students correctly answering each item are reported along with the mean of the difference.
- In this analysis each group is, in effect, acting as its own control.

Figure 14. Description of the observed, expected, and differences in strand and item performance strategy for a group.

The example of a district-level report used in the focus group was based on information generated from this strategy and is shown in Figure 15.

Strand	Item#	Item Description	District's Observed Percent Correct	District's Expected Percent Correct	Difference from Expected
Strand 1: Number & Operation	D1 - 08		42	36	6
	D1 - 15		76	78	-2
	D1 - 01		51	52	-1
	D1 - 09		72	70	2
	D1 - 07		28	24	4
	D1 - 06		55	52	3
	D1 - 05		34	25	9
	D1 - 03		69	69	0
	D1 - 02		90	95	-5
Mean			**57**	**55**	**2**
Strand 2: Algebra	D1 - 16		93	94	-1
	D1 - 04		58	59	-1
	D1 - 11		74	75	-1
	D1 - 12		72	75	-3
	D1 - 13		76	78	-2
	D1 - 14		66	64	2
	D1 - 10		55	55	0
Mean			**71**	**71**	**-1**
Strand 3: Geometry	D1 - 23		59	62	-3
	D1 - 17		58	63	-5
	D1 - 18		82	86	-4
	D1 - 32		63	64	-1
	D1 - 24		63	52	11
	D1 - 25		71	66	5
	D1 - 20		80	93	-13
	D1 - 34		56	51	5
	D1 - 21		48	53	-5
	D1 - 19		27	24	3
Mean			**61**	**61**	**-1**
Strand 4: Measurement	D1 - 29		27	28	-1
	D1 - 22		18	30	-12
	D1 - 30		24	18	6
	D1 - 28		56	60	-4
	D1 - 27		54	52	2
	D1 - 26		57	68	-11
Mean			**39**	**43**	**-3**
Stand 5: Data Analysis & Probability	D1 - 33		55	58	-3
	D1 - 36		19	16	3
	D1 - 35		58	38	20
	D1 - 31		34	31	3
Mean			**42**	**36**	**6**

Figure 15: Example of a report of the observed, expected, and differences in strand and item performance strategy for a district.

Focus Group Results: Observed, Expected, and Differences in Strand and Item Performance for a Group

The focus-group discussion of this reporting format included a number of questions along with the discussion of strengths, weaknesses, and recommendations. The questions raised by the focus-group participants included the following.

Questions

- *Would this be a good chart for small school districts or do numbers of students make a difference? Could this chart be used for individual class/teacher?*
- *When a district normally doesn't do well in testing, this might be good news when reporting to the board, but how would they understand "expected" and "observed" performance?*
- *Why bother? [Although framed as a question, the context suggests this was a negative rhetorical question.]*

Strengths

The positive comments recorded by the participants on their worksheets and subgroup summary forms in response to reporting the differences between observed and expected performance of groups of students at the item level include the following.

- *When a district normally doesn't do well in testing, this might be good news when reporting to the board, …*
- *Like it, good information.*
- *With a friendly item description, this document might have value for a teacher.*
- *Item description is key.*
- *Very helpful to teachers that specific item descriptions are included.*
- *Item description very important along with format description.*

The verbal descriptions of what the items are measuring emerges in the comments of the focus-group participants as being important and they are a major value of this reporting approach.

Weaknesses

The areas of weakness recorded by the participants on their individual worksheets and subgroup summary forms in response to reporting the difference between observed and expected performance of groups of students at the item level include the following.

- *Expected is not optimal.*
- *Will not push low performing districts and schools to improve.*
- *Built-in comfort level.*

- *Could lead to mediocrity.*
- *Built in expectations padding – "I scored as expected; why try to improve?"*
- *Confusing and not useful for instruction.*
- *Report can be misleading.*
- *Data confusing.*
- *Not as good as the next approach.*
- *District level too broad.*
- *Doesn't directly impact instruction.*
- *Doesn't relate to NCLB.*

A major concern that arose in the discussion is the possibility that schools and districts might interpret a level of achievement on each strand that meets the expectations as an adequate level of performance. In fact, a relatively low-achieving school that meets expectation on all the strands is still performing at a low level and should not be comfortable or feel that such attainment is optimal.

Suggestions

The suggestions offered by the focus group for reporting the difference between observed and expected performance of groups of students at the item level include the following.

- *Would like specific item description meaning response format.*
- *Item descriptions rather than curriculum codes from standards.*
- *Add indication of how strands are weighed.*
- *Put standard next to description.*
- *Need at school level.*
- *Need to be developed by class/school as well. Mean should be carried two decimal points, not rounded. Would like to see indication of 2 point items.*
- *If used include decimals to hundredths.*

The importance of item descriptions emerges again, as it did in the listing of strengths of this approach. The suggestion that such reports could be provided at the school level is also offered.

Discussion and Conclusions

The first comment by a participant during the open discussion of this reporting format was, "This is a formula for mediocrity." While a bit harsh, it expresses the concern that a school or district might have item-level performance in line with expectations but the performance could still be quite low. The possibility that such "on level" performance might lead to complacency was clearly a concern to the group. The addition of a column reporting the

percent of students statewide who answered each item correctly could provide a normative comparison. With the inclusion of statewide data, a district could see that students might be achieving as expected but are still performing below the state level. The importance of careful descriptions of what is being measured, seen in the focus-group discussion of other reporting formats, emerged again as being valued by the participants. This approach is not recommended for use on the PACT assessments, but the value of careful descriptions of what is being measured is noted.

B6. Observed, Expected, and Differences in Strand/Item Performance at the Achievement Level Cut Scores

Introduction

This final approach is similar to the previous approach except that the reference group is not each district's own performance but the performance expected by students at each cut score. The IRT ability at each cut score and the IRT item difficulties can be combined in the IRT model to estimate the probability that students with the ability at each cut score will correctly answer each item on the test. This probability can be viewed as the proportion of students at each cut score expected to answer the items correctly.

The proportion of students at each cut score expected to answer the item correctly is then compared to the observed proportion of students in a district who answer the items correctly. The differences between these expected and observed proportions are the proportions of district students who did more or less well than students at the respective cut scores. A generic example of the type of score report that reflects this approach is shown below and the actual example used in the focus group is shown as Figure 18.

Strand	Item Number	Item Description	District Percent Correct	State Expected Percent at			District Percent Above/Below at		
				Basic	Proficient	Advanced	Basic	Proficient	Advanced

Figure 16. Generic report format for Strategy 6.

This strategy can be interpreted as reporting how well a district is doing on each item relative to students at each achievement-level cut score. This would tell districts "how far they have to go" to reach each standard or how far above each standard their students perform.

To facilitate interpretation, the items are ordered by strand and strand-level means are reported. The level of detail in the item descriptions is critical. At minimum, the objective that each item measures should be identified. The more detail provided in the descriptions, the more useful this report might be.

Description and Explanation of the Reporting Strategy and Format

The information presented to the focus group for this strategy is shown as Figure 17.

6. GROUP OBSERVED PERFORMANCE ON EACH ITEM COMPARED TO EXPECTED PERFORMANCE ON THE ITEMS FOR STUDENTS AT EACH CUTSCORE

- Similar to the previous approach, except that the reference group is not each school's or district's own performance but the performance expected by students at each cut score.
- The observed proportion of students in a group correctly answering each item is calculated.
- Each cut score represents an ability or achievement measure on the underlying scale.
- The ability or achievement measure at each cut score is used to predict what proportion of students at each cut score is expected to correctly answer each item
- The difference between the observed proportion of students in a group correctly answering each item and proportion expected to answer the item correctly at each cut score is calculated and reported by strand.
- For the items on each strand, the means for the observed and expected proportions (at each cut score) of students correctly answering each item is reported along with the means of the difference.

Figure 17. Description of the observed, expected, and differences in the strand and item performance at the achievement-level cut score strategy

An example of a district-level report based on this approach that provides state-level comparison performance data is provided as Figure 18.

Strand	Item#	Item Description	District Observed Percent Correct	State Expected Percent at			District Percent Above/Below at		
				Basic	Proficient	Advanced	Basic	Proficient	Advanced
Strand 1: Number & Operations	D1 - 08		42	41	71	85	1	-29	-43
	D1 - 15		76	81	94	97	-5	-18	-21
	D1 - 01		51	58	83	92	-7	-32	-41
	D1 - 09		72	75	91	96	-3	-19	-24
	D1 - 07		28	27	57	75	1	-29	-47
	D1 - 06		55	58	83	92	-3	-28	-37
	D1 - 05		34	14	30	39	20	4	-5
	D1 - 03		69	74	91	96	-5	-22	-27
	D1 - 02		90	94	98	99	-4	-8	-9
Mean			**57**	**58**	**78**	**86**	**-1**	**-20**	**-28**
Strand 2: Algebra	D1 - 16		93	47	49	50	46	44	43
	D1 - 04		58	65	87	94	-7	-29	-36
	D1 - 11		74	79	93	97	-5	-19	-23
	D1 - 12		72	79	93	97	-7	-21	-25
	D1 - 13		76	81	94	97	-5	-18	-21
	D1 - 14		66	70	89	95	-4	-23	-29
	D1 - 10		55	61	85	93	-6	-30	-38
Mean			**71**	**69**	**84**	**89**	**2**	**-14**	**-18**
Strand 3: Geometry	D1 - 23		59	68	88	95	-9	-29	-36
	D1 - 17		58	69	89	95	-11	-31	-37
	D1 - 18		82	87	96	98	-5	-14	-16
	D1 - 32		63	70	89	95	-7	-26	-32
	D1 - 24		63	59	84	92	4	-21	-29
	D1 - 25		71	72	90	95	-1	-19	-24
	D1 - 20		80	92	98	99	-12	-18	-19
	D1 - 34		56	58	83	92	-2	-27	-36
	D1 - 21		48	59	84	92	-11	-36	-44
	D1 - 19		27	28	58	76	-1	-31	-49
Mean			**61**	**66**	**86**	**93**	**-5**	**-25**	**-32**
Strand 4: Measurement	D1 - 29		27	14	28	38	13	-1	-11
	D1 - 22		36	35	66	81	1	-30	-45
	D1 - 30		24	20	48	68	4	-24	-44
	D1 - 28		56	66	87	94	-10	-31	-38
	D1 - 27		54	58	83	92	-4	-29	-38
	D1 - 26		57	73	91	96	-16	-34	-39
Mean			**42**	**44**	**67**	**78**	**-2**	**-25**	**-36**
Stand 5: Data Analysis & Probability	D1 - 33		55	64	86	94	-9	-31	-39
	D1 - 36		19	18	43	64	1	-24	-45
	D1 - 35		58	38	46	48	20	12	10
	D1 - 31		34	36	67	82	-2	-33	-48
Mean			**42**	**39**	**61**	**72**	**3**	**-19**	**-30**

Figure 18. Example of a report based on the observed, expected, and differences in strand and item performance at the achievement-level cut scores.

Focus Group Results: Observed, Expected, and Differences in Strand/Item Performance at the Achievement Level Cut Scores

The focus-group discussion of this reporting format included a number of questions along with the discussion of strengths, weaknesses, and suggestions. The questions raised by the focus-group participants include the following.

Questions

- *What would an item description look like?*
- *Item descriptions will include what language, etc.?*
- *One issue: What would be in item description box?*

These questions speak to the issues of the nature of the item descriptions.

Strengths

The positive comments recorded by the participants on their worksheets and subgroup summary forms in response to reporting the difference between the observed item performance and the expected item performance at the achievement-level cut scores include the following.

- *Like it, good information.*
- *Could help identify areas that need more focused effort.*
- *Useful, breaks down data into Basic, Proficient, Advanced*
- *Helpful because it shows discrepancy between current performance and Proficient expectation.*
- *I think this would be helpful at the school level.*
- *Might be good for research staff in district office.*
- *I like this for district report.*

The focus-group participants saw a number of useful applications for this reporting format.

Weaknesses

The areas of weakness recorded by the participants on their individual worksheets and subgroup summary forms in response to reporting the observed item performance and the expected item performance at the achievement-level cut scores include the following.

- *Confusing. Not useful at classroom level.*
- *Classroom teachers may have difficulty interpreting this.*
- *Not sure of utility for classroom teachers.*

- *Too much information for teachers.*
- *Too much for teachers.*
- *For larger districts, this still doesn't pinpoint where you need to work. We have 3000 third graders, how do I determine where those 300 are that need to be pulled to Proficient?*
- *Too many numbers lose sight of pictures.*
- *I sincerely hope that this detailed summary will not be developed because it would distort conclusions and causes the worst consequence that we don't intend to have: test-driven instruction.*
- *Not suitable to show to parents.*

The major concern expressed in these comments involves the belief that classroom teachers would not be able to use this information because of the amount and complexity of the information and the density of the score report.

Suggestions

The suggestions offered by the focus group for reporting the difference between observed item performance and expected item performance at the achievement-level cut scores include the following.

- *The item descriptions must be as detailed as possible.*
- *The more detailed the information the more decisions the schools could make.*
- *Make items specific.*
- *Item description very important along with format description.*
- *Add teacher-friendly item description.*
- *Put standard next to description.*
- *Item description rather than curriculum codes.*
- *Item description is very valuable and important to have.*
- *Would be good to know if multiple choice or extended responses.*
- *Descriptions – MC or constructed response.*
- *Add format – MC versus Constructed Response.*
- *Also, [item] format is important (put that in columns showing kind of item it was).*
- *Add school column and maybe have different form for schools.*
- *It would be great to measure expected growth for individual students.*
- *However, putting scores for school/state/district on this form would be great.*
- *Similar to what we got in 1999, [format] could be improved with color.*
- *Show how strands are weighted. More readable display would be a good idea; graphs would be a plus. Like #1 (maybe at 90% probability).*

- *Add "school" column.*
- *Similar to '99 reports.*

Focus-group participants indicated very clearly that the value of this reporting format would depend directly on the level of detail in the item descriptions. The group suggested that, in addition to providing a clear description of the item's content, identifying the format of the item as multiple choice or open-ended would be useful.

Discussion and Conclusions

The response to this reporting format was dominated by questions and suggestions about the nature and level of detail in the item descriptions. The discussion of the group as a whole suggested stronger support for this approach than the written comments developed individually or in subgroups. However, the stronger support emerging from the discussion of the group as a whole was predicated on the use of detailed item descriptions. This approach should be advanced with the addition of more detailed item descriptions. Ordering the items by difficulty within a strand might be a useful modification of this approach.

C. Rating of Score Reporting Strategies

Focus-group participants were asked to rate each of the six reporting formats in terms of its usefulness at the school and district levels. They completed the rating scale twice – after the facilitator had described all of the procedures but before any group discussion and sharing took place and then again at the end of the entire focus-group process.

In completing the rating, participants were directed to consider the following questions:

- Will a school or school district find this information helpful?
- How could a school or school district use this information?
- Could this information be modified to be more informative or useful?
- How can this information be best presented?
- Might there be any problems in how this information is used?

Participants rated all of the item mapping/score reporting strategies on a scale of 1 through 4 using the following:

1 = No use to educators.
2 = Limited use to educators.
3 = Considerable use to educators.
4 = Very useful to educators.

The focus-group members were asked to complete the form, first from the perspective of the classroom or classroom teacher and then again from the perspective of a school district administrator or coordinator. The rating response portion of the form that was used is shown below. The actual rating form is provided in Appendix D.

Item Mapping/Reporting Strategy	Classroom Teacher				District Administrator/Coord.			
	N	L	C	V	N	L	C	V
1. Graphical Mapping	1	2	3	4	1	2	3	4
2. Narrative Description	1	2	3	4	1	2	3	4
3. Strand Level – Individual	1	2	3	4	1	2	3	4
4. Strand Level – Groups	1	2	3	4	1	2	3	4
5. Observed vs. Expected, Same Group	1	2	3	4	1	2	3	4
6. Observed vs. Expected at Cut Scores	1	2	3	4	1	2	3	4

Results

The mean ratings for the six reporting formats are shown in Table 13. These data are presented for discussion purposes to provide a different way to examine the focus groups' opinions about the six score reporting formats. Statistical tests of difference were intentionally not performed to avoid overinterpretation of the results based on 16 volunteer participants.

Table 13

Mean Ratings of the Reporting Strategies

Reporting Strategy	Before Discussions		After Discussions		Overall Mean (Rank)
	Teacher/ School	District	Teacher/ School	District	
1. Item Content Objective Mapping	3.00	2.94	2.82	2.94	2.93 (2)
2. Achievement Level Narrative	3.25	3.00	3.12	3.00	3.09 (1)
3. Strand Level - Individual	2.13	1.87	1.41	1.47	1.72 (6)
4. Strand Level – Groups	2.44	3.13	1.93	2.75	2.56 (4)
5. Observed, Expected – Same Group	2.13	2.67	2.19	2.47	2.37 (5)
6. Observed, Expected – Cut Scores	2.27	3.00	2.87	3.19	2.83 (3)
Mean	**2.54**	**2.77**	**2.39**	**2.64**	

Results of the ratings include the following findings.

- Strategy 2: The achievement performance level narrative approach received the highest mean rating overall. and this reporting strategy was generally rated higher than the other strategies before and after discussion.
- Strategy 1: The item content objective mapping approach received the next highest ratings.
- Strategy 6 was rated as considerably more useful to educators than Strategy 5. Strategy 6 involves reporting the differences between how group of students did on each item (observed performance) and the expected performance for students at the achievement-level cut scores. Strategy 5 involves the observed and expected performance of the same group with no comparison to other groups.
- The mean of Strategy 6 increased from the first to the second rating.
- Strategy 3: The strand achievement levels approach for individual students was rated as having no use or limited use to educators.

Conclusions

Taken as a whole, these ratings suggest that the focus-group participants viewed the various approaches as roughly falling into three categories. Taking some license with the rating-scale descriptors, these three categories can be characterized as:

- Useful or considerably useful to educators
  - Strategy 2 – Achievement Performance Level Narrative
  - Strategy 1 – Item Content Objective Mapping
  - Strategy 6 – Observed, Expected, and Differences in Strand and Item-level Performance at the Achievement Level Cut Scores
- Possibly useful to educators for some purposes
  - Strategy 4 – Strand Achievement Levels for Groups
  - Strategy 5 – Observed, Expected, and Differences in Strand and Item-level, Performance for the Same Group
- Limited usefulness or not useful to educators
  - Strategy 3 – Strand Achievement Levels for Individual Students

The placement of the six score reporting approaches into these ordered categories based on the participants' ratings seems quite consistent with the findings from the review of focus-group written records and discussions.

Section 5

Discussion and Conclusions

A. Introduction and Section Overview

The purpose of this project was to explore, develop, and evaluate various approaches to score reporting that can be used to present students' test scores in ways that are as informative and helpful as possible to students, parents, and educators. This project studied the features and formats of score reports that increase their value to educators for identifying students' strengths and weaknesses and for designing, monitoring, and adjusting instructional programs.

The project had two major components. First, there is a review of assessment-reporting research literature and practices and second, a field-based study of score reporting formats. A brief summary and discussion of each of these project components is presented below, followed by a discussion of several other related issues.

B. The Review of Score Report Research and Practice

The review of assessment-reporting research and practice revealed that many educators have difficulty interpreting score reports from large-scale assessment programs. The review identified a wide range of features in score reports that can be manipulated to make score reports more informative and user friendly. This summary examines two features of score reports: 1) basic content, and 2) format, language, and display features.

Basic Content of a Score Report

In general, the score report results should be related as closely and explicitly as possible to the content standards the assessment is designed to examine. It is valuable to report at the finest level of detail or smallest assessment unit for which reliable information can be presented. The finest level of detail would be the test item, then content clusters such as strands (e.g., subscales or subdomains), and then the total test.

While it is essential to report results in relation to content standards, it is critical to present results in relation to performance standards as well. There are numerous procedures for reporting scores in relation to performance cut scores and performance levels. The important feature is that a reader would have a way to know where the score is located relative to a performance-level cut score or interval.

Many practitioners found some form of normative information useful in understanding assessment results. Locating students in achievement levels and reporting percentages of students at these levels for a school, district, or state serves this purpose. Traditional norm-referenced reporting such as percentiles can also be considered.

The reliability or precision of all score results should be reported. Reliability is related to the level of reporting and as the level of reporting becomes smaller (e.g., moves from groups to individuals, from total test to strands and items), the reliability of individual scores becomes lower. Thus, at the smaller levels of reporting, it is increasingly unwise and misleading to report individual scores and more appropriate to report scores of groups of students. A performance-referenced report should include information about the precision or reliability of the classification.

In summary, score report results should:

- Be related to content standards as clearly and explicitly and as possible.
- Be reported in relation to performance standards.
- Include some form of normative information.
- Be reported at the finest level of detail for which reliable information can be provided.
- Include information about precision for all scores presented.

Score Report Format, Language, and Display Features

The physical format of score reports is critical, and it is difficult to summarize the wide variety of specific suggestions about the format of score reports that emerged from the reviews in this study. The general or overall "look" of the reports is an essential feature. The ease with which a reader can find the most important information and the actual print elements seem to influence readers' responses to score reports. In regard to general format features, score reports should:

- Be clean, as simple as possible, and uncluttered.
- Highlight important results in some way, e.g., (boxes, bold face).
- Use select print features such as font size, style, and spacing.

Score reports are a unique type of reading material for most adults because they often include numbers, tables of numbers, graphs, charts, and narrative elements. The reviews offer a number of suggestions for the use of numeric and visually presented information. A summary of some of the key recommendations includes:

- Avoid jargon that would not be familiar to the intended audience.
- Avoid statistical terms.
- Provide an explanation or glossary for any measurement terms used.
- Use simple and clear graphs, charts, and tables.
- Use text to explain graphs, charts, and tables.

C. The Study of Score Report Features and Formats

A study was conducted to design, develop, and evaluate different types of score reports as part of this project. Field-based educators offered guidance on critical features that score reports might contain. Six reporting formats were developed that reflected this advice and the information from the reviews of research and practice. Educators then reviewed the six score reporting formats, and qualitative and quantitative evaluation information was collected.

The six reporting formats are:

1. Item Content Objective Mapping.
2. Achievement Performance Level Narrative.
3. Strand Achievement Level for Individual Students.
4. Strand Achievement Level for Groups.
5. Observed, Expected, and Differences in Strand and Item Performance for a Group.
6. Observed, Expected, and Differences in Strand and Item Performance at the Achievement Level Cut Scores.

A brief summary of the evaluation of each of the six reporting formats is presented to show the connection between the guidelines and principles from the review of score reporting research and practice and the issues raised by the educators reviewing these score report formats. The strategies and summaries are presented in order from most to least useful, as evaluated by the focus group participants.

Strategy 2 – Achievement Performance Level Narrative

The evaluation data show that educators found the narrative descriptions of achievement levels the most useful reporting format they reviewed. This score reporting format has several features identified in the reviews of research and practices. First, it is content referenced and, in fact, content referenced at a fairly fine level of detail. The descriptions were based on a review of items that measure specific learning objectives, and the learning objectives are the finest level of content classification in the assessment program. Second, the achievement levels provided a normative interpretation because they are ordered categories. Third, the achievement levels are presented in purely written format with no tables, charts, or graphs. Interestingly, the reviewers recommended that the narrative format be deconstructed and the results be presented in the form of key bullets. This approach would seem to have the effect of making the information clearer and more concise and highlights the key results.

Strategy 1 – Item Content Objective Map

The focus-group reviewers saw the graphical mapping of content objectives as having value and potential. This format is also content referenced in that the plot symbols used are words

or phrases that reflect the content being measured. The reviewers recommended that the plot symbols contain more detailed descriptions of the content. The map has a normative feature in that the location of the achievement levels is also shown on the graph. The reviewers expressed concern that the graphical format might be difficult for teachers to understand, and interpretive guides and professional development activities would be required. The review of research and practice suggests that graphs should be simple and concise, and the graph used in this approach to score reporting is neither.

Strategy 4 – Strand Achievement Levels for Groups

The review of this score reporting strategy was mixed but generally positive. The group saw some value in these approaches but expressed concerns. Reporting strand-level achievement for groups of students was seen as useful for some general purposes. The strands reflect the content and the levels represent performances and are, in some senses, normative information. There did not seem to be enough specificity in this approach to be useful to classroom teachers.

Strategy 6 – Observed, Expected, and Differences in the Strand and Item Performance at the Achievement Level Cut Scores

The review of this score reporting strategy was mixed but generally positive. Reporting the item-level performance of a school or school district compared to how students at the achievement-level cut scores were expected to perform appealed to members of the focus group. The potential value of this approach was connected to how much detail was used in describing the item and strand content. This approach has several features suggested as beneficial in the review of research and practice. The item and strand descriptions reference the content; performance is reported at the finest level of detail, namely the test item; and the performance achievement levels used are those that offer a normative feature to the report. The physical format of the report, however, leaves much to be desired. It is busy, cluttered, and complex and would need some serious supporting materials and explanation. If the information in the report could be presented in a more straightforward fashion, this score report might be seen as more useful.

Strategy 3 – Strand Achievement Levels for Individuals

The review of this score reporting strategy was mixed but mostly negative. Participants liked the idea of a score report providing information about students' performance on each strand in relation to the achievement levels. Such an approach would provide content and performance-related information for each student. The major concern about reporting achievement-level performance for individual students on each strand was the unreliability of the scores at the strand level. A detailed analysis of strand-level results showed that over 96% of the pair-wise differences in students' performance on the strands were not significantly different from random variation. Thus, inferences about students' strengths and

weaknesses based on comparing performance on the strands is misleading in all but a very few cases.

Strategy 5 – Observed, Expected, and Differences in the Strand and Item Performance for a Group

The evaluation data from the practitioners indicated clearly that this reporting strategy did not provide useful information. Having item-level data was viewed as a positive feature of the report. Developing expectations for a group based on its own overall level of achievement, however, was seen as dangerously misleading since a school or district could meet its expectation and there would be no external referent to indicate that the achievement was still quite weak.

The results of the study show considerable consistency with the guidelines and principles found in the review of the score reporting research and practices.

D. Other Issues

Use of Focus Groups

There are several references in the review of score reporting research and practice to the use of focus groups to evaluate various score reports designed for different audiences. The use of parent, teacher, and community focus groups to pilot test score reports is strongly recommended. No matter how carefully one might attend to the guidelines and suggestions offered in the review of research and practice, having potential users evaluate the usefulness of the report seems an invaluable step in developing informative and effective score reports.

Interpretive Guides and Other Materials

This study and review did not consider interpretive guides and other forms of support materials. The design, development, and distribution of print and web-based supplementary resources seems an important part of comprehensive reporting systems and the interested reader might examine Goodman and Hambleton (2003) who review these materials in their study.

The Actual Use of Score Reports

Finally, as mentioned earlier, it is important to be clear that this review of score reporting research and the study conducted as part of this project did not address the question of how teachers, principals, parents, and others actually use score reports. The review reported what researchers and practitioners thought and said about various score reports, not what practitioners actually did with them. A different line of research might involve researchers visiting schools and school district offices to observe and interview students, teachers, districts personnel, and parents. The purpose of this line of research would be to describe how the information in various score reports is used in practice.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 508-600). Washington, DC: American Council of Education.

Cizek, G. J, (1996). Setting passing scores. *Educational Measurement: Issues and Practices*, *15*(2), 20-31.

Cizek, G. J. (Ed.) (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Coster, W., Ludlow, L., & Mancini, M. (1999). Using IRT variable maps to enrich understanding rehabilitation data. *Journal of Outcome Measurement* , *3*(2), 123-133.

Flanagan, J.C. Units, scores and norms. In E. F. Linquist (Ed). *Educational measurement*. Washington, D.C.: American Council on education, 1950

Forte Fast, E., Blank, R. K., Potts, A., & Williams, A. (2002). *A guide to effective accountability reporting*. Washington, DC: Council of Chief State School Officers.

Forte Fast, E. & Tucker, C. (2001, April) *Redesign of the student assessment reporting system in Connecticut*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Goodman, D.P. & Hambleton, R. K. (2003). *Student test score reports and interpretive guides: Review of current practices and suggestions for future research*. (Center for Educational Assessment Research Report No. 477). Amherst, MA: University of Massachusetts, School of Education.

Hambleton, R. K. (2002). How can we make NAEP and state test score reporting scales and reports more understandable? In R. W. Lissitz & W. D. Schafer (Eds.), *Assessment in educational reform* (pp. 192-205). Boston: Allyn & Bacon.

Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers? *Educational Measurement: Issues and Practice, 10*(4), 16-18.

Huynh, H. (1976). On the reliability of decisions in domain referenced testing. *Journal of Educational Measurement*. *13*, 253 – 264.

Huynh, H, (1978). Reliability of multiple classifications. Psychometrika, *43*, 317-325.

Huynh, H. (1998). On score location of binary and partial credit items and their application to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics*. *23(*1), 35-56.

Jaeger, R. (1998). *Reporting the results of the National Assessment of Educational Progress* (NVS NAEP Validity Studies). Washington, DC: American Institutes for Research.

Linacre, J. (1999), *WINSTEPS*, Chicago: MESA Press.

Livingston, S. A. & Zieky, M. J. (1982). *Passing scores: A manual for setting standards for educational and occupational tests.* Princeton, NJ: Educational Testing Services.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Mitzel, H. C., Lewis, D. M., Patz, R. J., and Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed). *Setting performance standards: Concepts, methods, and perspectives*. (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Associates.

National Education Goals Panel. (1998). *Talking about tests: An idea book for state leaders.* Washington, DC: U.S. Government Printing Office.

National Research Council. (2001). *NAEP reporting practices: Investigating district-level and market-basket reporting.* Washington, DC: National Academy Press.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 1111, 115 Stat. 1449-1452 (2002)

Snodgrass, D., & Salzman, J. A. (2002, April). *Creating the Rosetta stone: Deciphering the language of accountability to improve student performance.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Stone, M. H., Wright, B. D., and Stenner A. J. (1999). Mapping variables. *Journal of Outcome Measurement*, *3*(4), 308-322.

Tufte, E.R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Tufte, E.R. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.

Tukey, J. W. (1990). Data-based graphics: Visual display in the decades to come. *Statistical Science, 5*(3), 327-339.

Wainer, H. (1990). Graphical visions from William Playfair to John Tukey. *Statistical Science, 5*(3), 340-346.

Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher, 21*(1), 14-23.

Wainer, H. (1997a). Improving tabular displays: With NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics, 22*(1), 1-30.

Wainer, H. (1997b). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot.* New York, NY: Copernicus Books.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, *36*(4), 301-335.

Wainer, H., & Thissen, D. (1981). Graphical data analysis. In M. R. Rosenweig & L. W. Porter (Eds.), *Annual Review of Psychology* (pp. 191-241). Palo Alto, CA: Annual Reviews.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press

Wright, B. D. & Stone, M. (1979) *Best test design*. Chicago: MESA Press.

Ysseldyke, J., & Nelson, J. R. (2002). Reporting results of student performance on large-scale assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 467-480). Mahwah, NJ: Lawrence Erlbaum Associates.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 20*(2), 15-25.

Appendices

Appendix A.1

Focus Group 1 Participants

The Director of the Office of Assessment, in consultation with staff and others, selected educators for this meeting. Educators were selected who could represent different types of schools, regions of the state, constituencies, and educational perspectives. The participants and their affiliations are shown below.

Participant	Affiliation
Min Ching	Richland District One
Archie Franchini	Estill High School (Hampton Two)
Debra Hamm	Richland District Two
Judy Ingle	Georgetown District Office
Andrea Keim	SCDE Curriculum and Standards
Pat Mohr	SCDE Curriculum and Standards
Judy Newman	Sumter District Office
Carol Poole	Berkeley District Office
Tom Pritchard	Horry District Office
Christina Schneider	SCDE- NAEP
Missy Wall-Mitchell	Lexington/Richland Five

From the Office of Assessment

Teri Siskind
Necati Engec
Joe Saunders

Appendix A.2

Focus Group 2 Participants

The Director of the Office of Assessment, in consultation with staff and others, selected educators for this meeting. Educators were selected who could represent different types of schools, regions of the state, constituencies, and educational perspectives. The participants and their affiliations are shown below.

Participant	Affiliation
Min Ching	Richland District One
Jennifer Gouvin	Richland District Two
Debra Hamm*	Richland District Two
Kathy Howard	Spartanburg District Five
Judy Ingle	Georgetown District Office
James Ann Lynch	Lake Carolina Elementary Richland One
Pat Mohr	SCDE Curriculum and Standards
Judy Newman	Sumter District Office
Jane Pulling	Marion District Seven
Tom Pritchard	Horry District Office
Sue Rischell	Lake Murray Elem. Lexington/Richland Five
Janelle Rivers	Lexington District One
Cindy Saylor	SCDE Curriculum and Standards
Christina Schneider	SCDE - NAEP
Llewellyn Shealy	Hand Middle School, Richland One
Laura Timmons	Satchel Ford Elementary Richland One
Randall Wall	Beaufort Middle School
Missy Wall-Mitchell	Lexington/Richland Five
Wanda Whatley	Lexington District Three

* Was not present at the focus group but reviewed all focus group materials and responded to the focus group questions by means of an extensive phone interview with the researcher.

From the Office of Assessment

Teri Siskind
Necati Engec
Joe Saunders

Appendix B

Intercorrelations Among the Grade 3, Mathematics Strands

Subscale	2. Algebra	3. Geometry	4. Measurement	5. Data Analysis & Probability
1. Numeration/ Operations	0.62	0.60	0.61	0.52
2. Algebra		0.55	0.58	0.48
3. Geometry			0.58	0.50
4. Measurement				0.51

Appendix C

Classification Agreement and *Kappa* Indices for Grade 3, Mathematics Strands

Test/ Strand	Number of Points	KR21	Raw Agreement	Kappa Index
Total Test	40	0.86	0.69	0.52
Strand 1	10	0.65	0.59	0.35
Strand 2	8	0.55	0.54	0.28
Strand 3	10	0.54	0.48	0.23
Strand 4	7	0.58	0.59	0.31
Strand 5	5	0.37	0.43	0.15

Appendix D

Focus Group Advisory Committee for PACT Interpretation
March 21, 2003

Committee Member Rating Form

Committee Member: _____

Summary Comparison for all Item Mapping/ Score Reporting Strategies

Directions: On this sheet, please consider all six Item Mapping/ Score Reporting Strategies.

As before, the questions to consider are:

Will a school or school district find this information helpful?
How could a school or school district use this information?
Could this information be modified to be more informative or useful?
How can this information be best presented?
Might there be any problems in how this information is used?

Please rate the six Item Mapping/ Score Reporting Strategies on a scale of 1 through 4 where

1 = No use to educators
2 = Limited use to educators
3 = Considerable use to educators
4 = Very useful to educators

On the back of this sheet, please identify the strategy you think would be most useful and briefly explain why you think it would be so useful.

Item Mapping/ Reporting Strategy	Classroom/Teacher				District			
	N	L	C	V	N	L	C	V
1. Graphical Mapping	1	2	3	4	1	2	3	4
2. Narrative Description	1	2	3	4	1	2	3	4
3. Strand Level - Individual	1	2	3	4	1	2	3	4
4. Strand Level – Groups	1	2	3	4	1	2	3	4
5. Observed vs. Expected, Same Group	1	2	3	4	1	2	3	4
6. Observed vs. Expected at Cut Scores	1	2	3	4	1	2	3	4

Focus Group Ad Hoc Advisory Committee for PACT Interpretation

March 21, 2003

Committee Member Response Sheet

Committee Member: _____

Item Mapping/ Score Reporting Strategy: **Group Observed Performance on Each Item Compared to Expected Performance on Each Item for Students at each Cut Score**

[*Author's Note: a separate form was provided for each score reporting format*]

The questions for discussion are:

Will a school or school district find this information helpful?

How could a school or school district use this information?

Could this information be modified to be more informative or useful?

How can this information be best presented?

Might there be any problems in how this information is used?

Please record any suggestions or issues you have, or that emerge in discussion, that you think are the important for the Department of Education to consider. Use the back of this sheet as necessary.

Vita Summary for Joseph M. Ryan

Arizona State University West

jmryan@asu.edu, 602-543-3411

Joseph Ryan has been an assessment advisor and measurement consultant for more than 20 years. He has worked with schools, school districts, state departments of education, and many test developers. He is currently a member of the Technical Advisory Committees for the states of Arizona, Ohio, South Carolina, and Washington, and has worked as a technical advisor with more than a dozen states. His areas of technical expertise include scaling, equating, standard setting, and bias or DIF analyses. He is especially interested in reporting procedures and formats that provide instructionally useful information to students, teachers, parents, and others.

Dr. Ryan is a professor of educational measurement and evaluation at Arizona State University West where he teaches courses in measurement, statistics, and research. He is the Director of the ASU West Research Consulting Center and was interim Dean of the College of Education in 1999. He currently provides technical support to the Arizona Assessment Collaborative, a consortium of some 20 school districts coordinated by WestEd that have developed standards-referenced English and Spanish language assessments in reading and mathematics.

Before moving to Arizona in 1995, Dr. Ryan was on the faculty at the University of South Carolina beginning in 1974. He was a professor in the department of Educational Psychology where he taught measurement, psychometrics, statistics, and ethnographic methods. In 1990, Dr. Ryan founded the South Carolina Center for Excellence in the Assessment of Student Learning (CEASL). This K-16 statewide center focused on in-service assessment professional development for educators at all levels. The Center provided print and electronic resources and offered more than 70 professional development workshops annually.

Dr. Ryan received an AB in mathematics and M Ed in Educational Psychology from Boston College and a PhD in Measurement, Evaluation, and Statistical Analysis from the University of Chicago. He is the co-author of two books, seven chapters, and more than 100 articles and papers. His most recent publications include chapters on *Standard setting issues, strategies, and procedures for combining data from multiple-measures to classify student* and *Variation in achievement scores related to gender, item formats, and content area tested*, in G. Tindel and T. Haladyna (Eds.) (2002), Issues, research, and recommendations for large-scale assessment programs. Mahwah, NJ: Lawrence Earlbaum Associates